Genome Engineering Technology and Its Application in Mammalian Cells

Dissertation Advisors: Dr. George M. Church and Dr. Feng Zhang Le Cong

Genome Engineering Technology and Its Application in Mammalian Cells

Abstract

The advancement of high-throughput, large-scale biochemical, biophysical, and genetic technologies has enabled the generation of massive amounts of biological data and allowed us to synthesize various types of biomaterial for engineering purposes. This enabled improved observational methodologies for us to navigate and locate, with unprecedented resolution, the potential factors and connections that may contribute to biological and biomedical processes. Nonetheless, it leaves us with the increasing demand to validate these observations to elucidate the actual causal mechanisms in biology and medicine. Due to the lack of powerful and precise tools to manipulate biological systems in mammalian cells, these efforts have not been able to progress at an adequate pace.

This work aims to bridge the gap between data generation and experimental verification by developing molecular-resolution genome engineering technologies that can effect cell-specific genetic and epigenetic perturbation in mammalian cells. The development of these novel tools starts from harnessing aspects of two prominent families of microbial systems, the Transcription Activator-like Effectors (TALEs) and the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/Cas systems. The development of these technologies was executed through several lines of engineering efforts. When applied in higher eukaryotes, these tools provide researchers with new

reverse engineering instruments to directly probe relevant biological molecules and pathways that are observed from analysis of biological data. In combination with sensitive and accurate readout methods in mammalian systems, these technologies could together establish transformative means for modeling the causal relationships between genetic or epigenetic variances and human disease, while also serving as the first step towards the development of rational molecular therapies for complex human disorders and of synthetic biology applications for the research community.

Table of Contents

Acknowledgements

My graduate studies have been a fascinating journey, filled with all kinds of emotions and discoveries. I would like to express my sincere gratitude to my advisors Dr. George Church and Dr. Feng Zhang for their support, guidance, and mentorship. I was not a biology student by training at Tsinghua University, but rather was studying in an engineering department with focus on rules and standards, and it is they who introduced me to the world of scientific research. From both of my advisors, I observed, understood, and absorbed the essence of thinking critically and challenging the well-established.

I started my graduate work in late 2009, when I knew little about genomics or biotechnologies. I tried to explore this new ground as much as I could and I appreciate discussions with George on significant questions in this field. I met Feng as a rotation student when he was working to develop synthetic tools for applications in generating model systems. I was entranced by this area and started to work with him on this topic. To continue working with Feng, I moved to the Broad Institute in January 2011. This transition offered me an opportunity to participate in the set-up of a new laboratory and to be part of extensive brainstorming sessions about new ideas and projects. I owe a great deal to my advisors for my evolution as a scientific thinker and for making these opportunities possible. Most importantly, I thank Feng for the time he invested in me as my research advisor for my projects at the Broad Institute, working together side by side on the same bench to train me as a scientist.

Moreover, I thank my Preliminary Qualifying Exam (PQE) committee and Dissertation Advisory Committee (DAC) members for their kind suggestions and critical opinions on my research proposal and progress, including Dr. David Altshuler, Dr.

Constance Cepko, Dr. Patricia D'Amore, Dr. Kevin Eggan, Dr. Michael Greenberg, and Dr. Jagesh Shah. I would also like to thank all the lab members that I have worked with from the Zhang lab, the Church lab, administrative staffs from my programs and institutions, my collaborators and my friends at Harvard, MIT, the Broad Institute, and other parts of the world. The intellectual and personal interactions with all of you, this group of brilliant and innovative scientists and professionals, are important and influential elements to my training. This environment is no doubt the rich soil that nurtures the type of intensely cutting-edge, extraordinarily iconoclastic, and unbelievably impactful science that I, feeling humbled and fortunate, have been able to engage in first at Harvard and later at the Broad Institute. The collective Boston-Cambridge community is my favorite and to me the most charming neighborhood in the U.S., and for that I would say everyone I met here are part of my memorable life in this town.

Finally, I would like to thank my family, especially my parents, my mom Shufang Lu and my dad Litian Cong, who taught me and made me this person I am today, who has always been my strongest support with unconditional love, to whom I owe too much to even know it all, and my girlfriend, Yushan Jiang, who have been truly supportive through these years, encouraging me and believing in me, and my grandparents, Jinhua Gao, Yixing Cong, Fengying Li, Zhenguo Lu, who influenced and imprinted on me their characters and to whom I would like to dedicate my work to, and my cousins, Huan Cong, Xiao Cong, Rui Miao, Shan Lu, Teng Ren. Their love is the most precious thing. There are so many people that have helped me in this journey and I don’t think my words or the space here would be sufficient to thank all of them, but I want to say that I am always grateful and I wish you all the very best.

List of Tables and Figures

1. General Introduction

The emergence of high-throughput biology has generated tremendous amount of data, which in combination with enhancement of observational tools exemplified by super-high resolution microscopy, has greatly improved our observation of biological details related to human diseases. Almost every field of study from the classic genetics to relatively recent stem cell biology is going through an explosion of new hypotheses for various biological processes from molecular level to organismic level.

Ensuing efforts to systematically validate these hypotheses and quantitatively verify and apply these models require new powerful tools for these genome-scale investigations. New technologies centered around the specific perturbation of genome properties opened up the possibility for accurate and efficient genome-scale engineering, such as designer zinc-finger protein (ZFP), multiplex genome engineering and accelerated evolution (MAGE), complete whole-genome synthesis, among others (*1-8*). However, despite this advancement in biological manipulation techniques, our ability to perform this type of research is still limited by the lack of genome engineering tools that are precise, easy to design, quick to implement, low-cost and readily deliverable into mammalian cells (*1, 2, 5*).

With respect to earlier work in the field, zinc-finger nucleases have been synthesized by fusing the zinc finger DNA recognition module to a catalytic domain from nucleases such as FokI. This kind of chimeric nuclease have been shown to be functional to promote homologous recombination at target site, although off-target effects and toxicity are major issues that need to be addressed (*9, 10*). Over the years, several reports have confirmed the feasibility of designing a

customized zinc-finger protein using various approaches and demonstrated their functionality (*3, 4, 11-16*). Nonetheless, it is very difficult to apply this technology generally due to the unreliability of the zinc-finger binding domain in terms of its DNA-binding affinity and specificity, and the difficulty in scaling up the system in mammalian cells. To date, the most well-established method to build a customized zinc finger protein is to utilize randomized library screening with phage display technique, yeast two-hybrid library, or bacterial screening library (*10, 12, 14, 17-22*). All these issues have been underlying obstacles for the generalization and application of zinc-finger based technology, and the monopoly of a few commercial companies has prevented the wide application of this system in the broad scientific community.

Recent reports demonstrated the modularity of DNA-binding domain from Transcription Activator Like Effectors (TALEs), a group of microbial effector proteins from a diversity of plant bacteria *(23-26)*. My thesis work focuses on two different projects to harness this system to address the challenges for developing genome engineering technologies in mammalian cells. The first part of my work describes my effort to engineer programmable genome targeting tools based on TALEs to activate gene expression and introduce genome modifications (*27, 28*). The second part of my work focuses on the optimization and development of TALE technology to improve its targeting accuracy and expand additional functions, in particular, the ability to repress transcription at endogenous loci in mammalian systems (*29*).

Similarly, the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/Cas system in prokaryotes has been studied by several groups to elucidate their organizations and functions in bacteria cells (*30-32*). We decided to try to engineer this new family of biological system for mammalian genome engineering purposes. This third part of my research employs CRISPR/Cas

systems from prokaryotic cells to develop a multiplexable genome engineering system that is simple, efficient, fast to deploy, and cost-effective (*33*).

The contents of these three parts of work have been published in peer-reviewed research journals as cited above and in each chapter. Together the efforts of my research constitute a series of mammalian genome engineering technologies that have been designed and developed to address the lack of powerful, precise, affordable tools to meet a critical need for biological and biomedical research.

2. Efficient Construction of Sequence-specific TAL Effectors for Modulating Mammalian Transcription

This work is done with Dr. Feng Zhang as equal contributors, and this entire chapter has been published as Zhang F*[1], Cong L*, et al. Nature Biotechnology. 2011 (*27*).

2.1 Existing genome engineering tools and basic properties of TAL effectors

Systematic interrogation and engineering of biological systems in normal and pathological states depend on the ability to manipulate the genome of target cells with efficiency and precision(*1, 34*). Some naturally occurring DNA binding proteins have been engineered to enable sequence-specific DNA perturbation, including designer polydactyl zinc finger (ZFs)(*14, 17, 35*) and meganuclease(*36, 37*) proteins. In particular, designer ZFs can be attached to a wide variety of effector domains such as nucleases, transcription effectors, and epigenetic modifying enzymes to carry out site-specific modifications near their DNA binding site. However, due to the lack of a simple correspondence between amino acid sequence and DNA recognition, design and development of sequence-specific DNA binding proteins based on designer ZFs and meganucleases remain difficult and expensive, often involving elaborate screening procedures and long development time on the order of several weeks. Here we developed an alternative DNA targeting platform based on the naturally occurring transcription activator-like effectors (TALEs) from *Xanthomonas sp.(23-26)*

[1] * indicates equal contribution.

TALEs are natural effector proteins secreted by numerous species of *Xanthomonas* to modulate host gene expression and facilitate bacterial colonization and survival(*24, 26*). Recent studies of TALEs have revealed an elegant code linking the repetitive region of TALEs with its target DNA binding site(*23, 25*). Common among the entire family of TALEs is a highly conserved and repetitive region within the middle of the protein, consisting of tandem repeats of mostly 33 or 34 amino acid segments (Fig. 2.1a). Repeat monomers differ from each other mainly in amino acid positions 12 and 13 (variable diresidues), and recent computational and functional analysis*(24, 26)* have revealed a strong correlation between unique pairs of amino acids at positions 12 and 13 and the corresponding nucleotide in the TALE binding site (e.g. NI to A, HD to C, NG to T, and NN to G or A; Fig. 2.1a). The existence of this strong association suggests a potentially designable protein with sequence-specific DNA binding capabilities, and the possibility of applying designer TALEs to specify DNA binding in mammalian cells. However, our ability to test the modularity of the TALE DNA binding code remains limited due to the difficulty in constructing custom TALEs with specific tandem repeat monomers. Early studies have tested the DNA binding properties of TALEs(*23, 38-42*), including two studies that tested artificial TALEs with customized repeat regions(*39, 40*).

2.2 Efficient method for synthesis of the modular, repetitive, sequence-specific DNA binding domain of TAL effectors

A prerequisite for exploring the modularity TALE repeat monomers is the ability to synthesize designer TALEs with tailored repetitive DNA binding domains. While this has been recently shown to be possible(*38-40*), the repetitive nature of the TALE DNA binding domains renders routine construction of novel TALEs difficult when using PCR-based gene assembly or serial

DNA ligation, and may not be amenable to high-throughput TALE synthesis. Furthermore, even though commercial services can be employed for the synthesis of novel TALE binding domains(*40*), they present a cost-prohibitive option for large scale TALE construction and testing. Hence a more robust protocol to construct large numbers of designer TALEs would enable ready perturbation of any genome target in many organisms.

To enable high-throughput construction of designer TALEs, we developed a reliable hierarchical ligation-based strategy to overcome the difficulty of constructing TALE tandem repeat domains (Fig. 2.1b and Supplementary Methods). To reduce the repetitiveness of designer TALEs and to facilitate amplification using PCR, we first optimized the DNA sequence of the four repeat monomers (NI, HD, NN, NG) to minimize repetitiveness while preserving the amino acid sequence. In order to assemble the individual monomers in a specific order, we altered the DNA sequence at the junction between each pair of monomers, similar to the Golden Gate cloning strategy for multi-piece DNA ligation(*43, 44*). Using different codons to represent the junction between each pair of monomers (Gly-Leu), we designed unique 4 base pair sticky-end ligation adapters for each junction (Supplementary Methods). Using this strategy, 4 monomers can be ligated simultaneously to form 4-mer tandem repeats. Three 4-mer repeats can be simultaneously ligated to form the desired 12-mer tandem repeat and subsequently ligated into a backbone vector containing a 0.5 length repeat monomer specifying the 13th nucleotide of the binding site at the C-terminus of the repeat domain, as well as the N- and C-terminal non-repetitive regions from the *Xanthomonas campestris pv. armoraciae* TALE *hax3* (Fig. 2.1b). Using this method, we attempted to construct 17 artificial TALEs with specific combinations of 12.5-mer repeats to target 14 base pair DNA binding sites – TALEs require the first letter of the binding site to be a

T, and the 12.5-mer repeat targets a 13 bp binding site. We analyzed 2 clones for each of the 17 dTALEs via sequencing and found that all dTALEs were accurately assembled. Furthermore, we were able to construct 16 dTALEs in parallel in 3 days, a time substantially shorter than what is required for constructing a similar number of dTALEs via commercial DNA synthesis, and at a fraction of the cost.

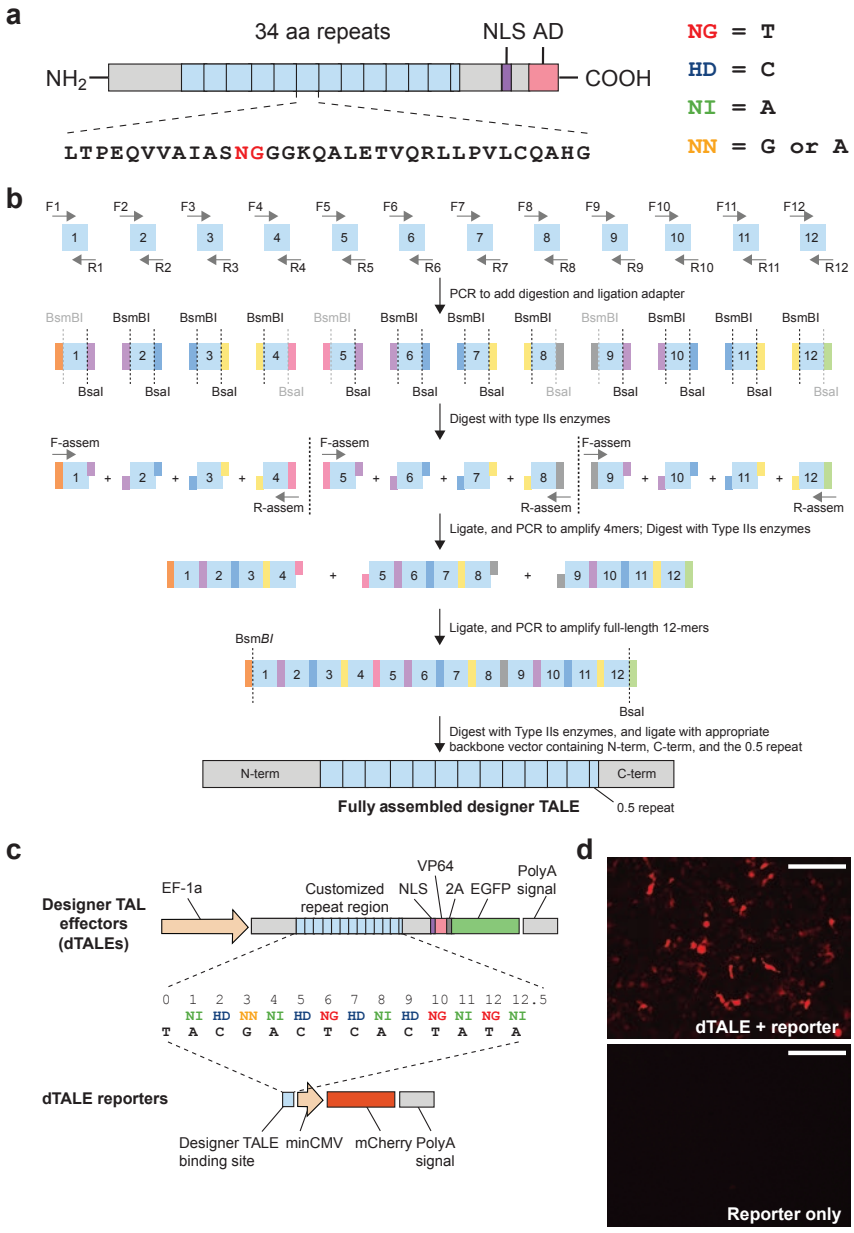


Figure 2.1 Design and construction of customized artificial transcription activator like effectors (dTALEs) for use in mammalian cells. a, Schematic representation of the native TALE hax3 from Xanthomonas campestris pv. armoraciae depicting the tandem repeat domain and the two variable repeat diresidues (red) within each repeat monomer that specify the base recognition specificity. The four most common naturally occurring diresidues that were used for

(Figure 2.1, continued) the construction of customized artificial TAL effectors are listed together with their proposed major base specificity. b, Schematic of the hierarchical ligation assembly method for the construction of customized dTALEs. 12 individual PCRs are performed for each of the 4 types of repeat monomers (NI, HD, NG, and NN) to generate a set of 48 monomers to serve as assembly starting material. Each of the 12 individual PCR products for a given monomer type (i.e. NI) has a unique linker specifying its programmed position in the assembly (color-coded digestion and ligation adapters). After enzymatic digestion with a Type IIs cutter (e.g. BsaI), unique overhangs (generated by leveraging the alternate codons for each amino acid in the junction) are generated. The unique overhangs facilitate the positioning of each monomer in the ligation product. The ligation product was PCR amplified subsequently to yield the full-length repeat regions, which were then cloned into a backbone plasmid containing the N- and C-termini of the wild type TALE hax3. c, Schematic representation of the fluorescence reporter system for testing dTALE-DNA recognition. The diagram illustrates the composition of the tandem repeat for a dTALE and its corresponding 14bp DNA binding target in the fluorescent reporter plasmid. NLS, nuclear localization signal; AD, activation domain of the native TAL effector; VP64, synthetic transcription activation domain; 2A, self-cleavage peptide. d, 293FT cells co-transfected with a dTALE plasmid and its corresponding reporter plasmid exhibited significant level of mCherry expression compared to the reporter-only control. Scale bar, 200μm.

2.3 Designer TALEs (dTALEs) effiiently targets desired DNA sequences

The DNA binding code of TALEs was identified based on analysis of TALE binding sites in plant genomes(*23, 25*) and the binding specificity of TALEs have been analyzed using various *in vitro* and *in vivo* methods(*23, 38, 40-42, 45-47*). In order to determine whether this code can be used to target DNA in mammalian cells, we designed a fluorescence-based reporter system (Supplementary Fig. 1) by placing the DNA binding site for each dTALE upstream of a minimal CMV promoter driving the fluorescence reporter gene mCherry (Fig. 2.1c). To generate dTALE

transcription factors, we replaced the endogenous nuclear localization signal (NLS) and acidic transcription activation domain (AD) of wild type *hax3* with a mammalian NLS derived from the simian virus 40 large T-antigen and the synthetic transcription activation domain VP64 (Fig. 2.1c). To allow quantitative comparison of the dTALE activity, we also fused a self-cleaving green fluorescent protein (GFP) to the C-terminus of each dTALE, so that we could quantify the relative level of dTALE expression using GFP fluorescence measurements.

Co-transfection of a dTALE (dTALE1) and its corresponding reporter plasmid in the human embryonic kidney cell line 293FT led to robust mCherry fluorescence (Fig. 2.1d). In contrast, transfection of 293FT cells with the reporter construct alone did not yield appreciable levels of fluorescence (Fig. 2.1d). Therefore, dTALEs are capable of recognizing their target DNA sequences, as predicted by the TALE DNA binding code, in mammalian cells. We quantified the level of reporter induction by measuring the ratio of total mCherry fluorescence intensity between cells co-transfected with dTALE and its corresponding reporter plasmid, and cells transfected with the reporter plasmid alone. To account for differences in dTALE expression level, we use the total GFP fluorescence from each dTALE transfection as a normalization factor to assess the fold of reporter induction fold.

Next we asked whether the DNA recognition code is sufficiently modular so that dTALEs could be customized to target any DNA sequence of interest. We first synthesized 13 distinct dTALEs targeting a range of DNA binding sites with diverse DNA sequence compositions (Fig. 2.2a) and found that 10 out of 13 dTALEs (77%) drove robust mCherry expression (> 10 folds) from their corresponding reporters. Three dTALEs exhibited more than 50 folds reporter induction

(dTALE1, dTALE4 and dTALE8), and only one out of 13 dTALEs (dTALE11) generated less than 5 fold induction of mCherry reporter expression (Fig. 2.2a). As a positive control, we constructed an artificial zinc finger-VP64 (ZF-VP64) fusion, where the ZF has previously been shown to activate transcription from a binding site in the human *erbB-2* promoter(*48*). This artificial ZF-VP64 protein was tested using the same mCherry reporter assay and demonstrated approximately 16-fold mCherry reporter activation (Fig. 2.2a).

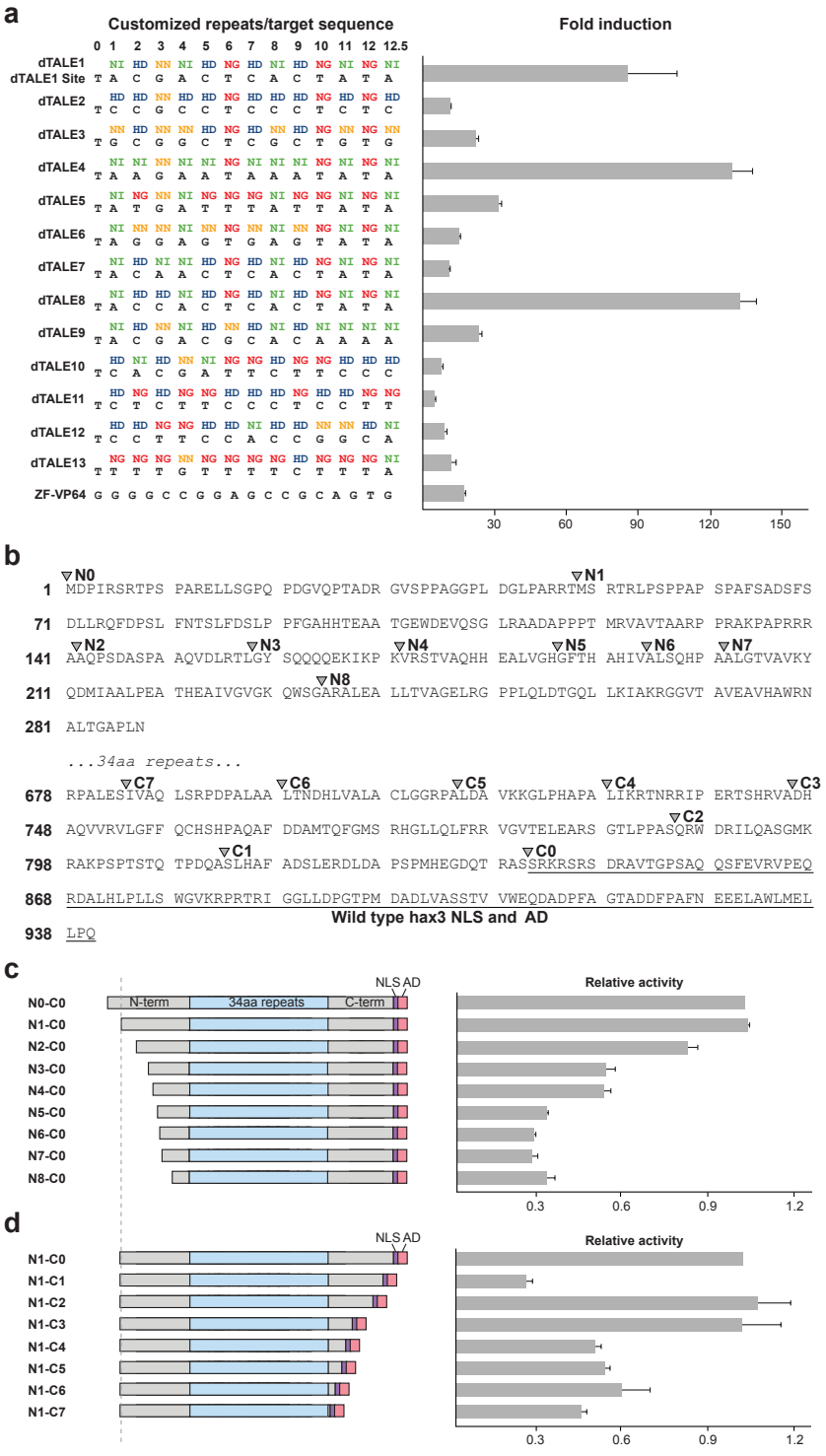


Figure 2.2 Functional characterization of the robustness of dTALE-DNA recognition in mammalian cells and truncation analysis of TALE N- and C-termini. a, 13 dTALEs were tested with their corresponding reporter constructs. Customized repeat regions and binding site

(Figure 2.2, continued) sequences are shown on the left. The activities of the dTALE to activate target gene expression are shown on the right as the fold induction of the mCherry reporter gene in a log scale. b, The N- and C-terminal amino acid sequence of wild type TAL effector hax3 showing the positions of all N- and C-terminal truncation constructs tested in 293FT cells. N0 to N8 designates N-terminal truncation positions (N0 retains the full-length N-terminus), and C0 to C7 designate C-terminal truncations. The amino acids representing the nuclear localization signal and the activation domain in the native HAX3 protein were underlined. c, Sequential truncation at the N-terminus of dTALE1 led to decreasing levels of reporter activity. Each truncation constructs is designated by its corresponding N-terminal and C-terminal truncation positions as indicated in panel b. Cartoon representations of all truncation constructs are shown on the left; the relative activity of each dTALE truncation construct compared to the dTALE(N0-C0) is demonstrated on the right. This relative activity is calculated from the fold induction of the reporter gene. d, Sequential truncation at the C-terminus of dTALE1 was used to characterize the optimal length of the C-terminus. Truncation position is designated in panel b. The relative activity of each truncation design compared to dTALE(N1,C0) is shown on the right. All error bars indicate s.e.m.; n=3. The fold induction was determined via flow cytometry analysis of mCherry expression in transfected 293FT cells, and calculated as the ratio of the total mCherry fluorescence intensity of cells transfected with and without the specified dTALE, normalized by the GFP fluorescence to control for transfection efficiency differences, as detailed in the Online Methods.

These data indicate that sequence-specific dTALEs can be designed and synthesized to target a wide spectrum of DNA binding sites at a similar or greater level as artificial ZF-VP64 transcription factors. While most dTALEs exhibited robust transcription activation in our reporter assay, the large range of observed activity suggests that other effects might contribute to dTALE DNA-targeting efficacy. Possible causes might include differences in DNA-interacting

capabilities such as binding strength of individual repeat types, context-dependence of monomer binding strength, or complexities of mammalian transcription processes(*47*).

To further characterize the robustness of dTALEs activities and their DNA binding specificity, we altered the target nucleotides in the binding sites of dTALE1 and dTALE13 to test the impact of mismatch position and number on dTALE activity. In general, we found that dTALE activity is inversely correlated with the number of mismatches (Supplementary Figs. 2 and 3). However, the specific dTALE recognition rules most likely depend on a combination of positional and contextual effects as well as the number of mismatchs, and need to be further characterized in greater detail.

2.4 Optimization of dTALE architecture through serial truncation testing

Each fully assembled dTALE has more than 800 amino acids. Therefore we sought to identify the minimal N- and C-terminal capping region necessary for DNA binding activity. We used Protean (LASERGENE) to predict the secondary structure of the TALE N- and C-termini and truncations were made at predicted loop regions. We first generated a series of N-terminal dTALE1 truncation mutants and found that transcriptional activity is inversely correlated with the N-terminus length (Fig. 2.2b,c). Deletion of 48 amino acids from the N-terminus (truncation mutant N1-C0, Fig. 2.2c) retained the same level of transcription activity as the full length N-term dTALE1, while deletion of 141 amino acids from the N-term (truncation mutant N2-C0, Fig. 2.2c) retained approximately 80% of transcription activity. Therefore given its full transcriptional activity, we chose to use truncation position N1 for all subsequent studies.

Similar truncation analysis in the C-terminus revealed that a critical element for DNA binding resides within the first 68 amino acids (Fig. 2.2b,d). Truncation mutant N1-C3 retained the same level of transcriptional activity as the full C-terminus, whereas truncation mutant N1-C4 reduces dTALE1 activity by more than 50% (Fig. 2.2d). Therefore in order to preserve the highest level of dTALE activity, approximately 68 amino acids of the C-terminus of *hax3* should be preserved.

2.5 Designer TALEs is capacle of modulating endogenous gene transcription in mammalian cells

The modularity of the TALE code is ideal for designing artificial transcription factors for transcriptional manipulation from the mammalian genome. In order to test whether dTALE could be used to modulate transcription of endogenous genes, we designed 4 additional dTALEs to directly activate transcription of *SOX2, KLF4* in human genome, and *c-Myc, Oct4* in mouse genome. dTALE binding sites were selected from the proximal 200bp promoter region of each gene (Fig. 2.3a). To assay the DNA binding activity of the 4 new dTALEs, we used the mCherry reporter assay as in previous experiments. three out of four dTALEs (*SOX2*-dTALE, *KLF4*-dTALE, and cMyc-dTALE) exhibited greater than 20-fold of mCherry reporter activation (Fig. 2.3a,b).

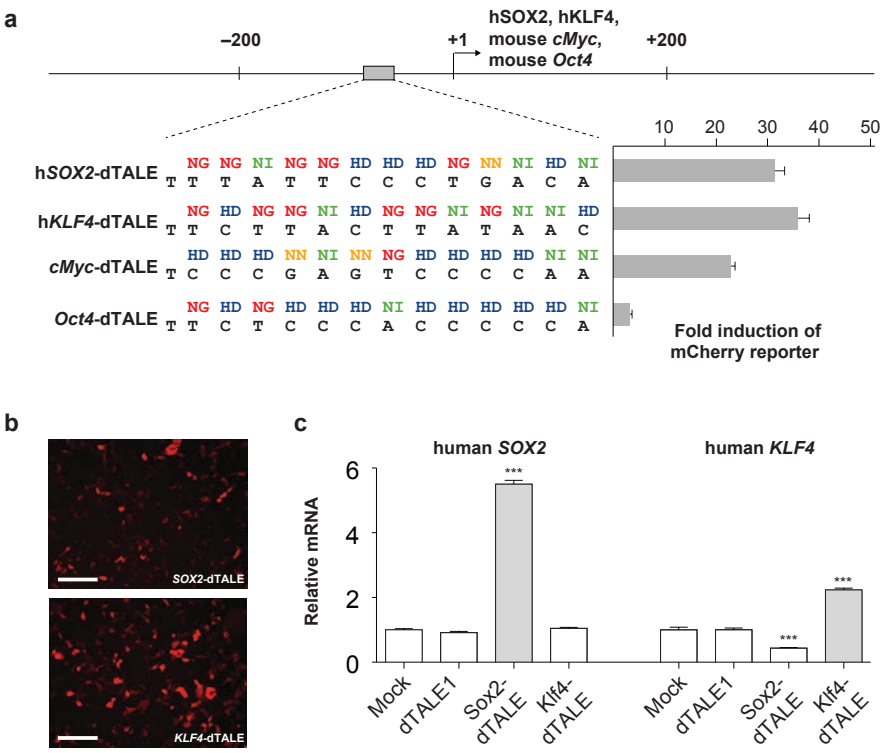


Figure 2.3 Activation of endogenous pluripotency factors from the human genome by designer TALEs. a, dTALEs designed to target the pluripotency factors, *SOX2*, *KLF4* in human geonme and *c-Myc*, *Oct4* in mouse genome, facilitate activation of mCherry reporter in 293FT cells. The target sites are selected from the 200bp proximal promoter region. The fold induction was determined via flow cytometry analysis using the same methodology as stated in Fig. 2.2 and detailed in Supplementary Methods. b, Images of dTALE induced mCherry reporter expression in 293FT cells. Scale bar, 200μm. c, mRNA levels of *SOX2* and *KLF4* in 293FT cells transfected with mock, dTALE1, *SOX2*-dTALE and *KLF4*-dTALE. Bars represent the levels of *SOX2* or *KLF4* mRNA in the transfected cell as determined via quantitative RT-PCR. Mock consists of cells receiving the transfection vehicle and dTALE1 is used as a negative control. The quantitation of the endogenous gene activation by dTALEs targeting the mouse genome is presented in *Supplementary Information*. All error bars indicate s.e.m.; n=3. *** $p < 0.005$.

To test the activity of dTALEs on endogenous genes, we transfected each dTALE targeting human or mouse genome into 293FT or Neuro2A cells respectively and quantified mRNA levels of each target gene using qRT-PCR. dTALE-*SOX2* and dTALE-*KLF4* were able to upregulate their respective target genes by 5.5 ± 0.1 and 2.2 ± 0.1 folds (Fig. 2.3c), providing a demonstration that dTALE can be used to modulate transcription from the genome. To control for specificity of activation, we transfected 293FT cells in parallel with dTALE1, which was not designed to target either *SOX2* or *KLF4*, and found no change in the level of Sox2 or Klf4 expression relative to the mock control. Interestingly, we observed a statistically significant decrease in the level of *KLF4* mRNA in 293FT cells transfected with *SOX2*-dTALE (approximately a 2-fold reduction). This is potentially due to secondary cross-regulation among reprogramming factors(*49, 50*). Similar activation of endognous gene expression is also observed in mouse Neuro2A cell lines with *cMyc*-dTALE and *Oct4*-dTALE (Supplementary Fig. 4). Finally, the degree of activation varies depending on the specific dTALE targets. This is not surprising as different genetic loci may not be equally accessible for activation, possibly due to epigenetic repression. Together, the data demonstrated that dTALE can be designed to bind and specifically activate transcription from the promoters of endogenous mammalian genes.

2.6 Implication and significance of TAL effector technology

The modular nature of the TALE DNA recognition code provides a novel and attractive solution for achieving sequence-specific DNA interaction in mammalian cells. For the first time, sequence-specific DNA binding proteins with predictable binding specificity can be generated in a matter of days, economically using molecular biology methods accessible to most. Future studies exploring the molecular basis of TALE-DNA interaction will likely extend the modular

nature of the TALE code for increased precision, specificity, and robustness. Given the ability of dTALEs to efficiently anchor transcription effector modules to endogenous genomic targets, other functional modules, including nucleases(*8*), recombinases(*51*), and epigenetic modifying enzymes(*34*), can be similarly targeted to specific binding sites. The designer TALE toolbox will empower researchers, clinicians, and technologists alike with a new repertoire of programmable precision genome engineering technologies.

2.7 Material and methods

2.7.1 Design and construction of designer TALEs and reporters

To simplify construction of designer TALEs, a dTALE backbone containing the N-term, a single 0.5 repeat regions carrying the variable diresidue NI, and C-term of *hax3* was synthesized (DNA2.0) and cloned into a lentiviral expression vector containing the mammalian ubiquitous EF-1α promoter (pLECYT)(*52*). To allow for insertion of customized repeat domains, a linker containing two Type IIs BsmBI sites are inserted between the N-term and the 0.5 repeat region. A DNA fragment containing a mammalian NLS, transcription activation domain VP64, and 2A-GFP was assembled via PCR assembly and fused to the C-term of the synthesized dTALE backbone (pLenti-EF1a-dTALE(0.5 NI)-WPRE). HD, NG, and NN versions of the backbone were generated via site directed PCR mutagenesis using QuikChange II XL (Stratagene). The full nucleotide sequences for the four backbone vectors are available in Supplementary Information. Customized dTALE repeat domains were synthesized via hierarchical ligation of individual repeat monomers (Supplementary Methods). To minimize repetitiveness of the final assembled tandem repeat domain, the DNA sequence for each type of repeat monomer (HD, NG, NI, or NN) has been optimized by altering the amino acid codons. The sequences for the

optimized monomers are listed in Supplementary Table 1, and the assembly primers are listed in Supplementary Table 2. mCherry reporter plasmids carrying dTALE binding site were generated by inserting sequences containing the binding site upstream of the minimal CMV promoter (Supplementary Fig. 1.)

2.7.2 Cell culture and reporter activation assay

The human embryonic kidney cell line 293 FT (Invitrogen) was maintained under 37ºC, 5% CO2 using Dulbecco's modified Eagle's Medium supplemented with 10% fetal bovine serum, 2mM GlutaMAX (Invitrogen), 100U/mL Penicillin, and 100µg/mL Streptomycin. mCherry reporter activation was tested by co-transfecting 293FT cells with plasmids carrying dTALEs and mCherry reporters. 293FT cells were seeded into 24- or 96-well plates the day prior to transfection at densities of $2x10^5$ cells/well or $0.8x10^4$ cells/well respectively. Approximately 24h after initial seeding, cells were transfected using Lipofectamine 2000 (Invitrogen). For 24-well plates we used 500ng of dTALE and 30ng of reporter plasmids per well. For 96-well plates we used 100ng of dTALE and 7ng of reporter plasmids per well. All transfection experiments were performed according to manufacturer's recommended protocol.

2.7.3 Flow cytometry

mCherry reporter activation was assayed via flow cytometry using a LSRFortessa cell analyzer (BD Biosciences). Cells were trypsinized from their culturing plates approximately 18 hours after transfection and resuspended in 200ul of media for flow cytometry analysis. The flow cytometry data was analyzed using BD FACSDiva (BD Biosciences). At least 25,000 events

were analyzed for each transfection sample. The fold induction of mCherry reporter gene by dTALEs was determined via flow cytometry analysis of mCherry expression in transfected 293FT cells, and calculated as the ratio of the total mCherry fluorescence intensity of cells from transfections with and without the specified dTALE. All fold induction values were normalized to the expression level of dTALE as determined by the total GFP fluorescence for each transfection.

2.7.4 Endogenous gene activation assay

293FT cells were seeded in 6 well plates. 4ug of dTALE plasmid was transfected using Lipofectamine 2000 (Invitrogen). Transfected cells were cultured at 37^{o}C for 48 hours, sorted for GFP positive population using BD FACSAria (BD Biosciences) to obtain cells that were successfully transfected and expressing dTALE. At least 1,000,000 cells were harvested and subsequently processed for total RNA extraction using the RNAeasy Mini Kit (Qiagen). cDNA was generated using the iScript cDNA Synthesis Kit (Bio-Rad) according to the manufacturer's recommended protocol. SOX2 and KLF4 mRNA were detected using TaqMan Gene Expression Assays (Applied Biosystems: Sox2 - Hs00602736_s1, Klf4 - Hs01034973_g1).

3. Methods and Protocols of A Transcription Activator-Like Effector Toolbox for Genome Engineering

The optimization and development of the methods and protocols are done with Drs. Neville Sanjana and Yang Zhou as equal contributors, and the work described in this chapter has been published as Sanjana NE*, Cong L*, Zhou Y*, et al. Nature Protocols. 2012 (*28*).

3.1 Introduction

Systematic reverse engineering of the functional architecture of the mammalian genome requires the ability to perform precise perturbations on gene sequences and transcription levels. Tools capable of facilitating targeted genome editing and transcription modulation are essential for elucidating the genetic and epigenetic basis of diverse biological functions and diseases. Recent discovery of the transcription activator-like effector (TALE) code(*23, 25*) has enabled the generation of custom TALE DNA binding domains with programmable specificity(*27, 39, 53-60*). When coupled to effector domains, customized TALEs provide a promising platform for achieving a wide variety of targeted genome manipulations(*27, 53, 54, 57, 60-62*). Previously, we reported efficient construction of TALEs with customized DNA binding domains for activating endogenous genes in the mammalian genome(*27*). Here we describe an improved protocol for rapid construction of customized TALEs and methods to apply these TALEs to achieve endogenous transcriptional activation(*27, 53, 54, 57*) and site-specific genome editing(*39, 53, 56, 58, 60-63*). Investigators should be able to use this protocol to construct TALEs for targets of their choice in less than one week.

3.1.1 Transcription Activator-Like Effectors.

TALEs are natural bacterial effector proteins used by *Xanthomonas sp.* to modulate gene transcription in host plants to facilitate bacterial colonization(*64, 65*). The central region of the protein contains tandem repeats of 34 amino acids sequences (termed monomers) that are required for DNA recognition and binding(*45, 46, 66, 67*) (**Fig. 3.1a**). Naturally occurring TALEs have been found to have a variable number of monomers, ranging from 1.5 to 33.5 (ref. 16). Although the sequence of each monomer is highly conserved, they differ primarily in two positions termed the repeat variable diresidues (RVDs, 12th and 13th positions). Recent reports have found that the identity of these two residues determines the nucleotide binding specificity of each TALE repeat and a simple cipher specifies the target base of each RVD (NI = A, HD = C, NG = T, NN = G or A)(*23, 25*). Thus, each monomer targets one nucleotide and the linear sequence of monomers in a TALE specifies the target DNA sequence in the 5` to 3` orientation. The natural TALE binding sites within plant genomes always begin with a thymine(*23, 25*), which is presumably specified by a cryptic signal within the non-repetitive N-terminus of TALEs. The tandem repeat DNA binding domain always ends with a half length repeat (0.5 repeat, **Fig. 3.1a**). Therefore, the length of DNA sequence being targeted is equal to the number of full repeat monomers plus two.

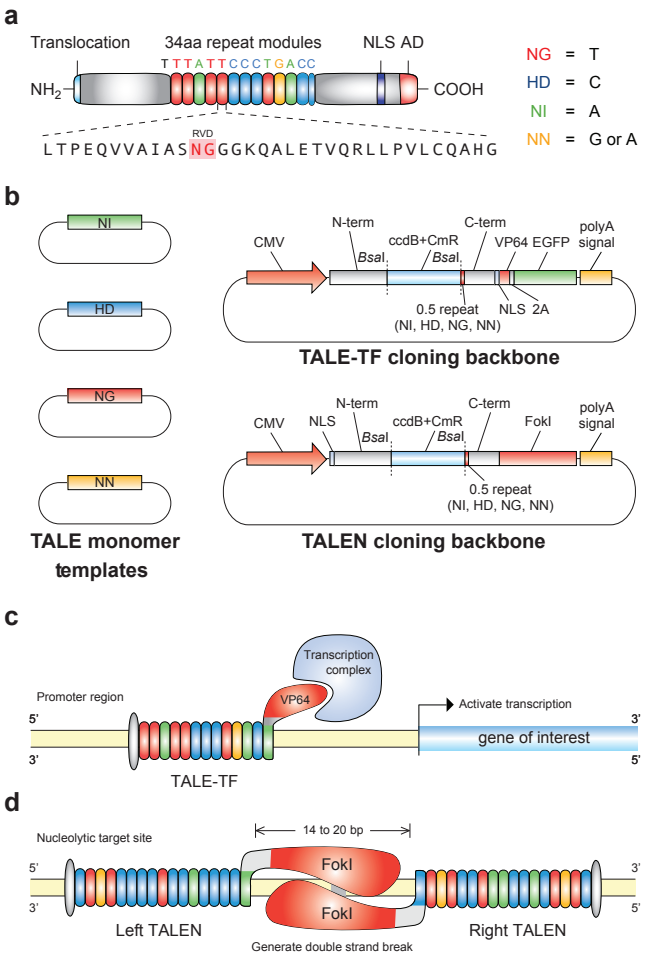


Figure 3.1 A TALE toolbox for genome engineering. (a) Natural structure of TALEs derived from *Xanthomonas sp.* Each DNA binding module consists of 34 amino acids, where the repeat variable diresidues (RVDs) in the 12th and 13th amino acid positions of each repeat specify the DNA base being targeted according to the cipher NG = T, HD = C, NI = A, and NN = G or A. The DNA binding modules are flanked by non-repetitive amino and carboxyl termini, which carry the translocation, nuclear localization (NLS), and transcription activation (AD) domains. A cryptic signal within the amino terminus specifies a thymine as the first base of the target site. (**b**) The TALE toolbox allows rapid and inexpensive construction of custom TALE-TFs and TALENs. The kit consists of 12 plasmids in total: 4 monomer plasmids to be used as templates for PCR amplification, 4 TALE-TF and 4 TALEN cloning backbones corresponding to 4 different bases targeted by the 0.5 repeat. CMV: cytomegalovirus promoter; N-term: non

(Figure 3.1, continued) repetitive amino terminus from the Hax3 TALE; C-term: non-repetitive carboxyl terminus from the Hax3 TALE; *Bsa*I: type IIs restriction sites used for the insertion of custom TALE DNA binding domains; ccdB+CmR: negative selection cassette containing the ccdB negative selection gene and chloramphenicol resistance gene; NLS: nuclear localization signal; VP64: synthetic transcriptional activator derived from VP16 protein of herpes simplex virus; 2A: 2A self-cleavage linker; EGFP: enhanced green fluorescent protein; polyA signal: polyadenylation signal; *Fok*I: catalytic domain from the *Fok*I endonuclease. (**c**) TALEs can be used to generate custom transcription factors (TALE-TFs) and modulate the transcription of endogenous genes from the genome. This schematic shows a TALE-TF designed to target the *SOX2* locus in the human genome. The *SOX2* TALE-TF recognizes the sense strand of the *SOX2* proximal promoter, and the recognition site begins with T. The TALE DNA-binding domain is fused to the synthetic VP64 transcriptional activator, which recruits RNA polymerase and other factors needed to initiate transcription. (**d**) TALE nucleases (TALENs) can be used to generate site-specific double strand breaks to facilitate genome editing through non-homologous repair or homology-directed repair. This schematic shows a pair of TALENs designed to target the AAVS1 locus in the human genome. Two TALENs target a pair of binding sites flanking a 16bp spacer. The left and right TALENs recognize the top and bottom strands of the target sites respectively. Each TALE DNA-binding domain is fused to the catalytic domain of *Fok*I endonuclease; when *Fok*I dimerizes, it cuts the DNA in the region between the left and right TALEN binding sites.

3.1.2 Comparison to other genome manipulation methods.

For targeted gene insertion and knockout, there are several techniques that have been used widely in the past, such as homologous gene targeting(*68-70*), transposases(*71, 72*), site-specific recombinases(*73*), meganucleases(*74*), and integrating viral vectors(*75, 76*). However most of

these tools target a preferred DNA sequence and cannot be easily engineered to function at non-canonical DNA target sites. The most promising, programmable DNA-binding domain has been the artificial zinc finger (ZF) technology, which enables arrays of ZF modules to be assembled into a tandem array and target novel DNA binding sites in the genome. Each finger module in a ZF array targets three DNA bases(*77, 78*). In comparison, TALE DNA binding monomers target single nucleotides and are much more modular than ZF modules. For instance, when two independent ZF modules are assembled into a new array, the resulting target site cannot be easily predicted based on the known binding sites for the individual finger modules. Perhaps the biggest caveat of ZFs is that most of the intellectual property surrounding the ZF technology platform is proprietary and expensive (>$10k per target site). A public effort for ZF technology development also exists through the Zinc Finger Consortium but the publicly available ZF modules can only target a subset of the 64 possible trinucleotide combinations(*27, 79, 80*). TALEs theoretically can target any sequence and have already been deployed in many organisms with impressive success (see Table 3.1). Although TALEs seem superior in many ways, zinc fingers have a much longer track record in DNA-targeting applications(*78*), including their use in human clinical trials(*81*). Despite their relatively recent development, early results with TALEs have been promising and it seems that they can be applied in the same way as zinc fingers for many DNA-targeting applications (e.g. transcriptional modulator(*27, 53, 54, 57*), nuclease(*39, 53, 56, 58, 60-63*), recombinase(*82-84*), transposase(*85, 86*)).

Table 3.1 Applications of custom TALEs on endogenous genome targets

	Species	Genomic Loci	References
TALE-TF	*A. thaliana*	EGL3	(*54*)
		KNAT1	
	H. sapiens	KLF4	(*27*)
		SOX2	
		NTF3	(*53*)
		PUMA	(*57*)
		IFNα1	
		IFNβ1	
TALEN	*S. cerevisiae*	URA3	(*58*)
		LYS2	
		ADE2	
	H. sapiens	CCR5	(*53*)
		NTF3	
		PPP1R12C (AAVS1)	(*61*)
		OCT4 (POU5F1)	
		PITX3	
	C. elegans	BEN-1	(*60*)
	D. rerio	HEY2	(*87, 88*)
		GRIA3A	
		TNIKB	
	R. norvegicus	IGM	(*89*)

3.1.3 Constructing customized TALE-TFs and TALENs.

Due to the repetitive nature of TALEs, construction of the DNA-binding monomers can be difficult. Previously, we and other groups have used a hierarchical ligation strategy to overcome the difficulty of assembling the monomers into ordered multimer arrays, taking advantage of degeneracy in codons surrounding the monomer junction and Type IIs restriction enzymes(*27, 55-59*). In this protocol, we employ the same basic strategy that we previously used(*27*) to construct TALE-TFs to modulate transcription of endogenous human genes. We have further improved the TALE assembly system with a few optimizations, including maximizing the dissimilarity of ligation adaptors to minimize misligations and combining separate digest and ligation steps into single Golden Gate(*43, 44, 90*) reactions. Briefly, we first amplify each nucleotide-specific monomer sequence with ligation adaptors that uniquely specify the monomer position within the TALE tandem repeats. Once this monomer library is produced, it can conveniently be re-used for the assembly of many TALEs. For each TALE desired, the appropriate monomers are first ligated into hexamers, which are then amplified via PCR. Then, a second Golden Gate digestion-ligation with the appropriate TALE cloning backbone (**Fig. 3.1b**) yields a fully-assembled, sequence-specific TALE. The backbone contains a *ccdB* negative selection cassette flanked by the TALE N- and C-termini, which is replaced by the tandem repeat DNA-binding domain when the TALE has been successfully constructed. *ccdB* selects against cells transformed with an empty backbone, therefore yielding clones with tandem repeats inserted(*56*).

Assemblies of monomeric DNA binding domains can be inserted into the appropriate TALE transcription factor (TALE-TF) or TALE nuclease (TALEN) cloning backbones to construct

customized TALE-TFs and TALENs. TALE-TFs are constructed by replacing the natural activation domain within the TALE C-term with the synthetic transcription activation domain VP64 (ref. 3) (**Fig. 3.1c**). By targeting a binding site upstream of the transcription start site, TALE-TFs recruit the transcription complex in a site-specific manner and initiate gene transcription. TALENs are constructed by fusing a C-term truncation (+63aa) of the TALE DNA binding domain(*53*) with the non-specific *Fok*I endonuclease catalytic domain (**Fig. 3.1d**). The +63aa C-term truncation has also been shown to function as the minimal C-term sufficient for transcriptional modulation(*27*). TALENs form dimers through binding to two target sequences separated by ~17 bases. Between the pair of binding sites, the *Fok*I catalytic domains dimerize and function as molecular scissors by introducing double-strand breaks (DSBs) (**Fig. 3.1d**). Normally, DSBs are repaired by the non-homologous end-joining(*91*) (NHEJ) pathway, resulting in small deletions and functional gene knock-out. Alternatively, TALEN-mediated DSBs can stimulate homologous recombination, enabling site-specific insertion of an exogenous donor DNA template(*53, 61*).

We also present a short procedure for verifying correct TALE assembly: using colony PCR to verify the correct insert length followed by DNA sequencing. With our cloning procedure, we routinely achieve high efficiency (correct length) and high accuracy (correct sequence). The cloning procedure is modular in several ways: We can construct TALEs to target DNA sequences of different lengths and the protocol is the same for producing either TALE-TFs or TALENs. The backbone vectors can be modified with different promoters to achieve cell-type specific expression.

Our protocol includes functional assays for evaluating TALE-TF and TALEN activity in human cells. This step is important because we have observed some variability in TALE activity on the endogenous genome, possibly due to epigenetic repression and/or inaccessible chromatin at certain loci. For TALE-TFs, we perform quantitative reverse-transcription polymerase chain reaction (qRT-PCR) to quantify changes in gene expression. For TALENs, we use the Surveyor mutation detection assay (i.e. the base-mismatch cleaving endonuclease *Cel2*) to quantify NHEJ. Although these assays are standard and have already been described elsewhere(*92, 93*), we feel that the functional characterization is integral to TALE production and therefore have presented it here with the assembly procedure. Other functional assays such as plasmid-based reporter constructs(*27, 56*), restriction sites destroyed by NHEJ(*94*), or other enzymes that detect DNA mismatch(*95*) may also be used to validate TALE activity.

Our protocol (**Fig. 3.2**) begins with the generation of a monomer library, which takes one day and can be re-used for building many TALEs. Using the monomer library, several TALEs can be constructed in a single day with an additional two days for transformation and sequence verification. To assess TALE function on the endogenous genome, we take ~3 days to go from mammalian cell transfection to qRT-PCR or Surveyor results.

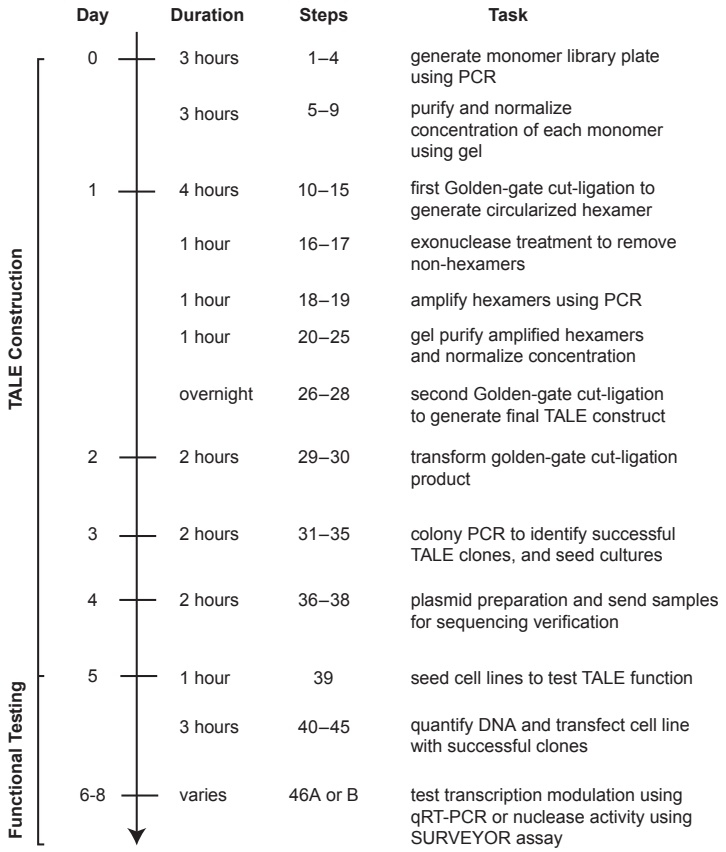


Figure 3.2 Timeline for the construction of TALE-TFs and TALENs. Steps for the construction and functional testing of TALE-TFs and TALENs are outlined. TALEs can be constructed and sequence verified in 5 days following a series of ligation and amplification steps. During the construction phase, samples can be stored at -20^{o}C at the end of each step and continued at a later date. After TALE construction, functional validation via qRT-PCR (for TALE-TFs) and Surveyor nuclease assay (for TALENs) can be completed in 2-3 days.

3.1.4 Comparison with other TALE assembly procedures.

A number of TALE assembly procedures have described the use of Golden-Gate cloning to construct customized TALE DNA binding domains(*27, 55-59*). These methods rely on the use of a large collection of plasmids (typically over 50 plasmids) encoding repeat monomers and intermediate cloning vectors. Our PCR-based approach requires significantly less initial plasmid

preparation, as our monomer library can be amplified on one 96-well PCR plate, and facilitates more rapid construction of custom TALEs. Plasmid-based amplification has a much lower mutation/error rate but, in our experience, the combination of a high-fidelity polymerase and the short length of the monomer template (~100 nt) results in accurate assembly. For building similar length TALEs to those presented in this protocol, the plasmid-based approaches also require an additional transformation and colony selection that extends the time needed to build TALEs. Thus, these alternative assembly protocols require a greater time investment both upfront (for monomer library preparation) and on a recurring basis (for each new TALE). For laboratories seeking to produce TALEs quickly, our protocol requires only a few hours to prepare a complete monomer library and less than a day to proceed from monomers to the final transformation into bacteria.

3.1.5 Targeting limitations.

There are a few important limitations with the TALE technology. Although the RVD cipher is known, it is still not well understood why different TALEs designed according to the same cipher act on their target sites in the native genome with different levels of activity. It is possible that there are yet unknown sequence dependencies for efficient binding or site-specific constraints (e.g. chromatin state) that are responsible for differences in functional activity. Therefore we suggest constructing at least 2 or 3 TALE-TFs or TALEN pairs for each target locus. Also, it is possible that engineered TALEs can have off-target effects – binding unintended genomic loci – which can be difficult to detect without additional functional assays at these loci. Given the relatively early state of TALE technology development, these issues remain to be addressed in a conclusive manner.

3.1.6 Experimental design.

TALE-TF target site selection. The programmable nature of TALEs allows for a virtually arbitrary selection of target DNA binding sites. As previously reported, the N-terminus of the TALE requires that the target site begin with a thymine nucleotide. For TALE-TFs, we have been successful targeting 14 to 20 bp sequences within 200bp of the transcription start site (**Fig. 3.1c**). It can be advantageous to select a longer sequence to reduce off-target activation, as it is known from reporter activation assays that TALEs interact less efficiently with targets contain more than one mismatching base. In our assembly protocol, we describe ligation of 18 monomers into a backbone containing a nucleotide-specific final 0.5 monomer; combined with the initial thymine requirement, this yields a total sequence specificity of 20 nucleotides. Specifically, the TALE-TF binding site takes the form 5`-$\mathtt{TN}^{19}$-3`. When selecting TALE-TF targeting sites for modulating endogenous gene transcription, we recommend selecting multiple target sites within the proximal promoter region (can target either the sense or antisense strand), as epigenetic and local chromatin dynamics might impede TALE binding. Larger TALEs might be beneficial for TALE-TFs targeting genes with less unique regions upstream of their transcription start site.

TALEN target site selection. Since TALENs function as dimers, a pair of TALENs, referred to as the left and right TALENs, need to be designed to target a given site in the genome. The left and right TALENs target sequences on opposite strands of DNA (**Fig. 3.1d**). As with TALE-TF, we design each TALEN to target a 20 base pair sequence. TALENs are engineered as a fusion of the TALE DNA-binding domain and a monomeric *Fok*I catalytic domain. To facilitate *Fok*I dimerization, the left and right TALEN target sites are chosen with a spacing of ~14-20 bases.

Therefore, for a pair of TALENs, each targeting 20 base pair sequences, the complete target site should have the form 5`-TN$^{19}$N$^{14\text{-}20}$N$^{19}$A-3`, where the left TALEN targets 5`-TN$^{19}$-3` and the right TALEN targets the antisense strand of 5`-N$^{19}$A-3` (N = A, G, T, or C). TALENs should have fewer off-target effects due to the dimerization requirement for the *Fok*I nuclease, although no significant off-target effects have been observed in limited sequencing verifications(*61*). Because DSB formation only occurs if the spacer between the left and right TALEN binding sites (**Fig. 3.1d**) is ~14-20 bases, nuclease activity is restricted to genomic sites with both the specific sequences of the left TALEN and the right TALEN with this small range of spacing distances between those sites. These constraints should greatly reduce potential off-target effects.

TALE monomer design. To ensure that all synthesized TALEs are transcribed at a similar level, all of the monomers have been optimized to share identical DNA sequences except in the variable di-residues – and are codon-optimized for expression in human cells (see Supplementary Data 3.1). This should minimize any difference in translation due to codon availability.

Construction strategy. Synthesis of monomeric TALE DNA binding domains in a precise order is challenging due to their highly repetitive nature. Previously(*27*), we took advantage of codon redundancy at the junctions between neighboring monomers and devised a hierarchical ligation strategy to construct ordered assemblies of multiple monomers. In this protocol, we describe a similar strategy but with several important improvements that make the procedure easier, more flexible, and more reliable (**Fig. 3.3**).

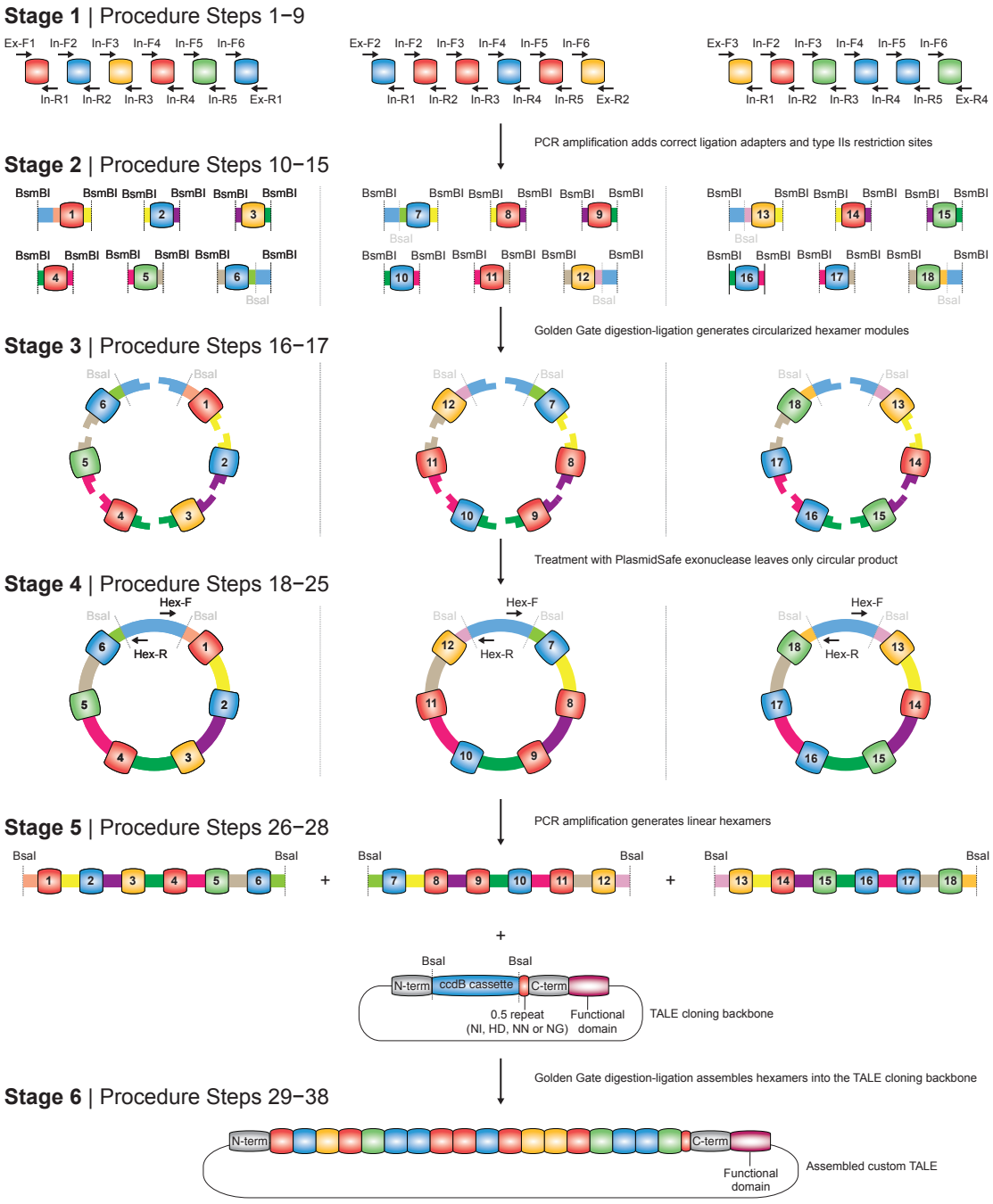


Figure 3.3 Construction of TALE DNA binding domains using hierarchical ligation assembly. Schematic of the construction process for a custom TALE containing a 18-mer tandem repeat DNA binding domain. Stage 1: specific primers are used to amplify each monomer and add the appropriate ligation adapters (Procedure Steps 1-9). Stage 2: hexameric tandem repeats

(Figure 3.3, continued) (1—6, 7—12, and 13—18) are assembled first using Golden Gate digestion-ligation. The 5` ends of monomers 1, 7, and 13 and the 3` ends of monomers 6, 12, and 18 are designed so that each tandem hexamer assembles into an intact circle (Procedure Steps 10-15). Stage 3: the Golden Gate reaction is treated with an exonuclease to remove all linear DNA, leaving only the properly assembled tandem hexamer (Procedure Steps 16-17). Stage 4: each tandem hexamer is amplified individually using PCR and purified (Procedure Steps 18-25). Stage 5: tandem hexamers corresponding to 1—6, 7—12, and 13—18 are ligated into the appropriate TALE-TF or TALEN cloning backbone using Golden Gate cut-ligation (Procedure Steps 26-28). Stage 6: The assembled TALE-TF or TALEN is transformed into competent cells and successful clones are isolated and sequence verified (Procedure Steps 29-38).

In our initial protocol(*27*), the digestion and ligation steps were carried out separately with an intervening DNA purification step. This improved protocol adopts the powerful Golden Gate cloning technique(*43, 44, 90*), requiring less hands-on time and resulting in a more efficient reaction. The Golden Gate procedure involves combining the restriction enzyme and ligase together in a single reaction with a mutually compatible buffer. The reaction is cycled between optimal temperatures for digestion and ligation. Golden Gate digestion-ligation capitalizes on Type IIs restriction enzymes, for which the recognition sequence is spatially separated from where the cut is made. During a Golden Gate reaction, the correctly ligated products no longer contain restriction enzyme recognition sites and cannot be further digested. In this manner, Golden Gate drives the reaction toward the correct ligation product as the number of cycles of digestion and ligation increases.

For the hierarchical ligation steps, we have optimized our previous cloning strategy for faster TALE production. The improved design takes advantage of a circularization step that allows

only properly assembled hexameric intermediates to be preserved (**Fig. 3.3**). Correctly ligated hexamers consist of six monomers ligated together in a closed circle, and incomplete ligation products are left as linear DNA. After this ligation step, an exonuclease degrades all non-circular DNA, leaving intact only the complete circular hexamers. Without circularization and exonuclease treatment, the correct ligation product would need to be gel purified before proceeding. The combination of Golden Gate digestion-ligation and circularization reduces the overall hands-on time required for TALE assembly.

Primer design for monomer library preparation. Each monomer in the tandem repeat must have its position uniquely specified. The monomer primers are designed to add ligation adaptors that enforce this positioning. Our protocol uses a hierarchical ligation strategy: For the 18mer tandem repeat, we first ligate monomers into hexamers. Then, we ligate three hexamers together to form the 18mer. By breaking down the assembly into two steps, we do not need unique ligation junctions for each monomer in the 18mer. Instead the same set of ligation junctions *internal* to each hexamer are re-used in all three hexamers (first ligation step), whereas unique (*external*) ligation junctions are used to flank each hexamer (second ligation step). As shown in **Fig. 3.4**, the internal primers used to amplify the monomers within each hexamer are the same, but the external primers differ between the hexamers. By re-using the same internal primers between different hexamers, our protocol minimizes the number of primers necessary for monomer amplification.

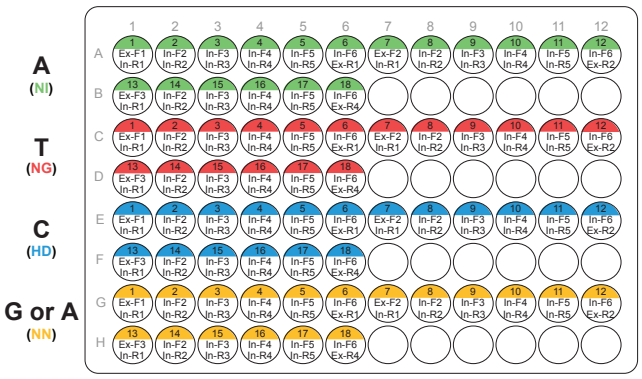


Figure 3.4 PCR plate setup used to generate a plate of monomers for constructing custom 18-mer TALE DNA binding domains. One 96-well plate can be used to carry out 72 reactions (18 for each monomer template). The position of each monomer and the primers used for the position is indicated in the well. Color coding in the well indicates the monomer used as the PCR template. Typically, 2-4 plates of 100 ul PCR reactions are pooled together and purified to generate a monomer library of sufficient quantity for production of many TALEs. During TALE construction, the corresponding monomer for each DNA base in the 18 bp target sequence can be easily picked from the plate.

Controls. As a negative control for Golden Gate assembly, we recommend performing a separate reaction with only the TALE-TF or TALEN backbone. Transformation of this negative control should result in few or no colonies due to the omission of the tandem repeats and resulting re-ligation of the toxic *ccdB* insert. After completing the TALE cloning, we use colony PCR or restriction digests to screen for correct length clones. For the final verification of proper assembly, we sequence the entire length of the tandem repeats. Due to limits in Sanger sequencing read length, other TALE assembly protocols have difficulty sequencing the entire tandem repeat region(*56, 58, 59*). The similarity of the monomers within the region makes primer annealing to specific monomers impossible. We have overcome this problem by slightly modifying the codon usage at the 5` end of monomer 7 to create a unique annealing site so that a

TALE with a 18mer DNA binding array can be verified through a combination of three staggered sequencing reads. Specifically, during the monomer amplification, the codons for the first 5 amino acids in monomer 7 are mutated via PCR to use different but synonymous codons, creating a unique priming site without changing the encoded TALE protein. This modification allows each hexamer in the 18mer to be sequenced with a separate sequencing read and requires only a standard read length of ~700 bp for complete sequence verification. For TALEs containing more than 18 full monomer repeats, we introduced a third unique priming site for sequencing at the 3` end of the 18th monomer using a similar approach. For construction of TALEs containing up to 24 full monomers with the entire tandem repeat region easily sequenced, see Box 3.1.

Design of functional validation assays. For TALE-TFs, qRT-PCR quantitatively measures the increase in transcription driven by the TALE-TF. For TALENs, the Surveyor assay provides a functional validation of TALEN cutting and quantifies the cutting efficiency of a particular pair of TALENs. These assays should be performed in the same cell type as intended for the TALE application, as TALE efficacy can vary between cell types, presumably due to differences in chromatin state or epigenetic modifications.

For qRT-PCR, we use commercially-available probes to measure increased transcription of the TALE-TF-targeted gene. For most genes in the human or mouse genomes, specific probes can be purchased (e.g. TaqMan Gene Expression Probes from Applied Biosystems). There are a wide variety of qRT-PCR protocols and, although we describe one of them here, others can be substituted. For example, a more economical option is to design custom, transcript-specific

primers (e.g. with NCBI Primer-BLAST) and use a standard fluorescent dye to detect amplified dsDNA (e.g. SYBR Green).

For Surveyor, we follow the recommendations given by the assay manufacturer when designing specific primers for genomic PCR. We typically design primers that are ~30 nucleotides long and with melting temperatures of ~65 °C. The primers should flank the TALEN target site and generate an amplicon of ~300-800 bp with the TALEN target site near the middle. During the design, we also check to make sure the primers are specific over the intended genome using NCBI Primer-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). Before using the primers for Surveyor, the primers and specific PCR cycling parameters should be tested to ensure that amplification results in a single clean band. In difficult cases where a single band product cannot be achieved, it is acceptable to gel extract the correct length band before proceeding with heteroduplex re-annealing and Surveyor nuclease digest.

3.2 Materials

3.2.1 Reagents

TALE construction

- TALE monomer template plasmids:

  pNI_v2

  pNG_v2

  pNN_v2

  pHD_v2

- TALE transcriptional activator (TALE-TF) plasmids:

  pTALE-TF_v2 (NI)

  pTALE-TF_v2 (NG)

  pTALE-TF_v2 (NN)

  pTALE-TF_v2 (HD)

- TALE nuclease (TALEN) backbone plasmids:

  pTALEN_v2 (NI)

  pTALEN_v2 (NG)

  pTALEN_v2 (NN)

  pTALEN_v2 (HD)

  These plasmids can be obtained individually or bundled together as a single kit from the Zhang Lab plasmid collection at Addgene (http://www.addgene.org/TALE_Toolbox). See Supplementary Data 1 for plasmid sequences.

- PCR primers for TALE construction (Table 3.3, Integrated DNA Technologies, custom DNA oligonucleotides)

- Herculase II Fusion polymerase (Agilent Technologies, cat. no. 600679)

CRITICAL Standard *Taq* polymerase, which lacks 3'-5' exonuclease proofreading activity, has lower fidelity and can lead to errors in the final assembled TALE. Herculase II is a high-fidelity polymerase (equivalent fidelity to *Pfu*) that produces high yields of PCR product with minimal optimization. Other high-fidelity polymerases may be substituted.

- 5x Herculase II reaction buffer (Agilent Technologies, included with polymerase)

- *Taq-B* polymerase (Enzymatics, cat. no. P725L)

- 10x *Taq-B* buffer (Enzymatics, included with polymerase)

- 25mM (each) dNTP solution mix (Enzymatics, cat. no. N205L)
- MinElute Gel Extraction Kit (Qiagen, cat. no. 28606)

CRITICAL MinElute columns should be stored at 4°C until use.

- QIAprep Spin Miniprep Kit (Qiagen, cat. no. 27106)
- QIAquick 96 PCR Purification (Qiagen, cat. no. 28181)
- UltraPure DNase/RNase-Free Distilled Water (Invitrogen, cat. no. 10977-023)
- UltraPure 10X TBE Buffer (Invitrogen, cat. no. 15581-028)
- SeaKem LE agarose (Lonza, cat. no. 50004)
- 10,000x SYBR Safe DNA stain (Invitrogen, cat. no. S33102)
- Low DNA Mass Ladder (Invitrogen, cat. no. 10068-013)
- 1 kb Plus DNA Ladder (Invitrogen, cat. no. 10787-018)
- TrackIt™ Cyan/Orange Loading Buffer (Invitrogen, cat. no. 10482-028)
- Restriction enzymes:
    - *BsmBI* (*Esp3I*) (Fermentas/ThermoScientific cat. no. ER0451)
    - *BsaI*-HF (New England Biolabs, cat. no. R3535L)
    - *AfeI* (New England Biolabs, cat. no. R0652S)
- Fermentas Tango Buffer and 10x NEBuffer 4 (included with enzymes)
- 100x Bovine Serum Albumin (New England Biolabs, included with *Bsa*I-HF)
- DL-Dithiothreitol (DTT) (Fermentas/ThermoScientific cat. no. R0862)
- T7 DNA ligase, 3,000 U/ul (Enzymatics, cat. no. L602L)

CRITICAL Do not substitute the more commonly-used T4 ligase. T7 ligase has 1000-fold higher activity on sticky ends than blunt ends and higher overall activity than commercially available concentrated T4 ligases.

- 10 mM Adenosine 5'-Triphosphate (New England Biolabs, cat. no. P0756S)
- Plasmid-Safe™ ATP-Dependent DNase (Epicentre, cat. no. E3101K)
- One Shot® Stbl3™ Chemically Competent *E. coli* (Invitrogen, cat. no. C7373-03)
- SOC medium (New England Biolabs, cat. no. B9020S)
- LB medium (Sigma, cat. no. L3022)
- LB agar medium (Sigma, cat. no. L2897)
- 100 mg/ml ampicillin, sterile-filtered (Sigma, cat. no. A5354)

TALEN and TALE-TF functional validation in mammalian cells

- HEK293FT cells (Invitrogen, cat. no. R700-07)
- Dulbecco's Minimum Eagle Medium (DMEM) (1X), high glucose (Invitrogen, cat. no. 10313-039)
- Dulbecco's Phosphate Buffered Saline (DPBS) (1X) (Invitrogen, cat. no. 14190-250)
- Fetal bovine serum, qualified and heat inactivated (Invitrogen, cat. no. 10438-034)
- Opti-MEM® I reduced-serum medium (Invitrogen, cat. no. 11058-021)
- GlutaMAX™-I (100X) (Invitrogen, cat. no. 35050079)
- Penicillin-streptomycin (100X) (Invitrogen, cat. no. 15140-163)
- Trypsin, 0.05% (1X) with EDTA•4Na (Invitrogen, cat. no. 25300-062)
- Lipofectamine 2000 ^{TM} transfection reagent (Invitrogen, cat. no. 11668027)
- QuickExtract ^{TM} DNA extraction solution (Epicentre, cat. no. QE09050)
- Herculase II Fusion polymerase (Agilent Technologies, cat. no. 600679)

CRITICAL Since Surveyor assay is sensitive to single-base mismatches, it is important to use only a high-fidelity polymerase. Other high-fidelity polymerases can be substituted; refer to the Surveyor manual for PCR buffer compatibility details.

- 5x Herculase II reaction buffer (Agilent Technologies, included with polymerase)
- Surveyor Mutation Detection Kit for Standard Gel Electrophoresis (Transgenomic, cat. no. 706025)

CRITICAL The Surveyor assay includes the *Cel2* base-mismatch nuclease. Alternatives include the *Cel1*, T7, mung bean, and S1 nucleases(*96, 97*). Of these, *Cel1* has been applied extensively for mutation detection(*98-100*) and established protocols are available for its purification(*98, 100*).

- Primers for Surveyor assay of TALEN cutting efficiency (see Experimental design for further information on Primer design, , Integrated DNA Technologies, custom DNA oligonucleotides**)**
- RNeasy Mini Kit (Qiagen, cat. no. 74104)
- QIAshredder (Qiagen, cat. no. 79654)
- 2-mercaptoethanol (Sigma, cat. no. 63689)

  ! CAUTION Wear appropriate personal protective equipment and work in a fume hood when handling 2-mercaptoethanol, which is acutely toxic and corrosive.
- RNAseZAP (Applied Biosystems, cat. no. AM9780)
- iScript cDNA synthesis kit (BioRad, cat. no. 170-8890)
- TaqMan® Universal Master Mix (Applied Biosystems, cat. no. 4364341)

- TaqMan® Gene Expression Assay Probes for the TALE-TF-targeted gene (Applied Biosystems, http://bioinfo.appliedbiosystems.com/genomic-products/gene-expression.html)

3.2.2 Equipment

- 96-well thermocycler with programmable temperature stepping functionality (Applied Biosystems Veriti, cat no. 4375786)

CRITICAL Programmable temperature stepping is needed for the TALEN (Surveyor) functional assay. Other steps only require a PCR-capable thermocycler.

- 96-well qPCR system (Applied Biosystems StepOnePlus™ Real-Time PCR System, Cat. No. 4376600)
- 96-well optical plates (Applied Biosystems MicroAmp, cat. no. N801-0560)
- 96-well PCR plates (Axygen, cat. no. PCR-96-FS-C)
- 8-well strip PCR tubes (Applied Biosystems, cat. no. N801-0580)
- QIAvac 96 vacuum manifold (Qiagen, cat. no. 19504)
- Gel electrophoresis system (BioRad PowerPac Basic Power Supply, cat no. 164-5050, and BioRad Sub-Cell GT System gel tray, cat. no. 170-4401).
- Digital gel imaging system (BioRad GelDoc EZ, cat. no. 170-8270, and BioRad Blue Sample Tray, cat. no. 170-8273)
- Blue light transilluminator and orange filter goggles (Invitrogen SafeImager 2.0, cat. no. G6600)
- Sterile 20 ul pipette tips for colony picking

- Gel quantification software (BioRad ImageLab, included with GelDoc EZ, or open-source NIH ImageJ, available at http://rsbweb.nih.gov/ij/)
- TALE online sequence verification software (Zhang Lab: http://taleffectors.com/tools/)
- 60 mm x 15 mm petri dishes (BD Biosciences, cat. no. 351007)
- Incubator for bacteria plates (Quincy Lab Inc, cat. no. 12-140E)
- Shaking incubator for bacteria suspension culture (Infors HT Ecotron)
- 6-well, cell culture-treated polystyrene plates (Corning, cat. no. 3506)
- UV spectrophotometer (ThermoScientific, cat. no. NanoDrop 2000c)

3.2.3 Reagent setup

Tris-borate EDTA (TBE) electrophoresis solution. Dilute in distilled water to 1x working solution for casting agarose gels and as a buffer for gel electrophoresis. Buffer can be stored at room temperature indefinitely.

10X BSA. Dilute 100x bovine serum albumin (BSA, supplied with *BsaI*-HF) to 10x concentration and store at -20 ^{o}C for at least 1 year in 20 ul aliquots.

10mM ATP. Divide 10mM ATP into 50ul aliquots and store at -20 ^{o}C for up to 1 year; avoid repeated freeze-thaw cycles.

10 mM DTT. Prepare 10 mM DTT solution in distilled water and store 20ul aliquots at -70 ^{o}C for up to 2 years; for each reaction, use a new aliquot since DTT is easily oxidized.

D10 culture medium. For culture of HEK293FT cells, prepare D10 culture medium by supplementing DMEM with 1X GlutaMAX and 10% fetal bovine serum. As indicated in the protocol, this medium can also be supplemented with 1X penicillin-streptomycin. D10 medium can be made in advance and stored at 4 ^{o}C for up to 1 month.

3.3 Procedure

3.3.1 Amplification and normalization of monomer library with ligation adaptors for 18mer TALE DNA binding domain construction

TIMING 6 hr

1| *Prepare diluted forward and reverse monomer primer mixes.* In a 96-well PCR plate, prepare primer mixes for amplifying a TALE monomer library (**Figure 3.3**, stage 1). Mix forward and reverse primers for each of the 18 positions according to the first two rows (A and B) of **Figure 3.4** and achieve a final concentration of 10uM for each primer. If using multi-channel pipettes, arrange the oligonucleotide primers in the order indicated in **Figure 3.4** to allow for easy pipetting. Typically, prepare 50ul mixes for each primer pair (40ul ddH_2O, 5ul 100uM forward primer, 5ul 100uM reverse primer).

2| Set up two 96-well monomer library plates following the organization shown in **Figure 3.4**; each plate will contain a total of 72 PCR reactions (18 positions for each monomer × 4 types of monomers). Although it is acceptable to have smaller volume PCR reactions, we typically make the monomer set in larger quantities since one monomer library plate can be used repeatedly for the construction of many TALEs. Each PCR reaction should be made up as follows to a total volume of 200ul, and then split between the two 96-well plates so that each well contains a 100ul PCR reaction:

Component	**Amount**	**Final concentration**

Monomer template plasmid (5 ng/ul)	2 ul	50 pg/ul
100mM dNTP (25mM each)	2 ul	1 mM
5X Herculase II PCR buffer	40 ul	1x
20uM primer mix (10 uM forward primer and 10 uM reverse primers from Step 1)	4 ul	200 nM
Herculase II Fusion polymerase	2 ul	
Distilled water	150 ul	
Total	**200 ul (for 2 reactions)**	

3| Perform PCR on the reactions from Step 2 using the following cycling conditions:

Cycle number	**Denature**	**Anneal**	**Extend**
1	95°C, 2 min		
2-31	95°C, 20 s	60°C, 20 s	72°C, 10 s
32			72°C, 3 min

4| After the reaction has completed, use gel electrophoresis to verify that monomer amplification was successful. Cast a 2% agarose gel in 1x TBE electrophoresis buffer with 1x SYBR Safe dye. The gel should have enough lanes to run out 2 ul of each PCR product from Step 3. Run the gel at 15 V/cm for 20 minutes. It is not necessary to check all 72 reactions at this step; it is sufficient to check all 18 reactions for one type of monomer template. Successful amplification should show a ~100 bp product. Monomers

positioned at the ends of each hexamer (monomers 1, 6, 7, 12, 13 and 18) should be slightly longer than the other monomers due to the length difference of the longer external primers.

? TROUBLESHOOTING

5| Pool both of the 100 ul PCR plates into a single deep-well plate. Purify the combined reactions using the QIAquick 96 PCR Purification kit following the manufacturer's directions. Elute the DNA from each well using 100 ul of Buffer EB (included with kit) pre-warmed to 55°C. Alternatively, PCR products can also be purified using individual columns found in standard PCR cleanup kits.

◆ CRITICAL STEP Before eluting the DNA, let the 96-well column plate air dry, preferably at 37°C, for 30 minutes on a clean Kimwipe so that all residual ethanol has enough time to evaporate.

6| *Normalization of monomer concentration.* Cast a 2% agarose gel. The gel should have enough lanes to run out 2 ul of each purified PCR product from Step 5. Include in one lane 10 ul of the quantitative DNA ladder. Run the gel at 20 V/cm for 20 minutes.

7| Image the gel using a quantitative gel imaging system. Monomers 1, 6, 7, 12, 13, and 18 are ~170 bp in size, whereas the other monomers are ~150 bp size (**Fig. 3.5a**, lanes 1-6). Make sure the exposure is short enough so that none of the bands are saturated.

8| Quantify the integrated intensity of each PCR product band using ImageJ or other gel quantification software. Use the quantitative ladder with known concentrations (5, 10, 20, 40, 100 ng) to generate a linear fit and quantify the concentration of each purified PCR product.

? TROUBLESHOOTING

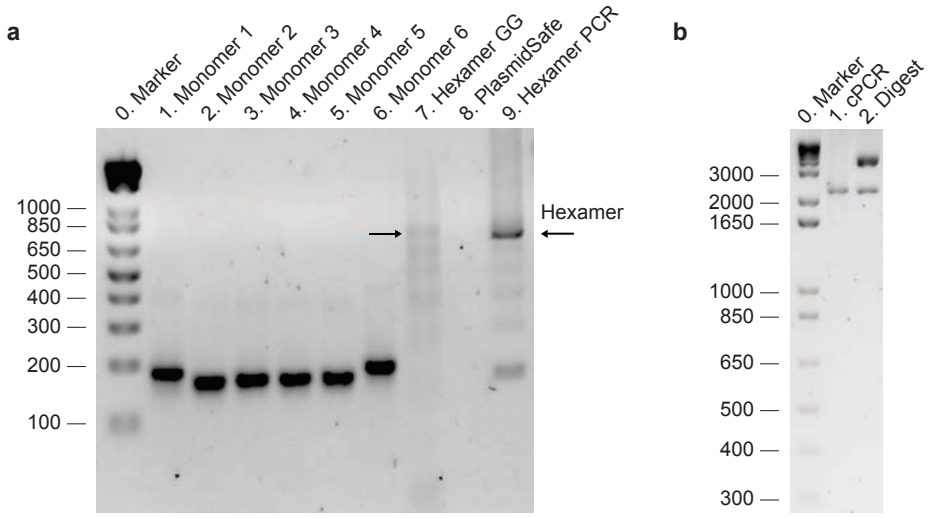


Figure 3.5 Example gel results from the TALE construction procedure. (a) Lanes 1—6: products from the monomer PCR reaction (Stage 1 in **Figure 3**) after purification and gel normalization (Procedure Steps 8-9). The molar concentrations of samples shown on this gel have been normalized so that equal moles of monomers are mixed for downstream steps. Monomers 1 and 6 are slightly longer than monomers 2, 3, 4, and 5 due to the addition of sequences used for circularization. Lane 7: result of the hexamer Golden Gate cut-ligation (Procedure Step 15). A series of bands with size ~700 bp and lower can be seen. Successful hexamer Golden Gate assembly should show a band ~700 bp (as indicated by arrow). Lane 8: hexamer assembly after PlasmidSafe exonuclease treatment (Procedure Step 17). Typically the amount of circular DNA remaining is difficult to visualize by gel. Lane 9: result of hexamer amplification (Procedure Step 20). A ~700 bp band should be clearly visible. The hexamer gel

(Figure 3.5, continued) band should be gel-purified to remove shorter DNA fragments. (**b**) Properly assembled TALE-TFs and TALENs can be verified using bacterial colony PCR (2175 bp band, lane 1) (Procedure Step 35) and restriction digest with *AfeI* (2118 bp band for correctly assembled 18-mer in either backbone; other bands for TALE-TF are 165 bp, 3435 bp, 3544 bp; other bands for TALEN are 165 bp, 2803 bp, 3236 bp; digest shown is for TALE-TF backbone vector, lane 2) (Procedure Step 35).

9| Adjust the plate of purified PCR products by adding Buffer EB so that each monomer has the same molar concentration. Since monomers 1, 6, 7, 12, 13, and 18 are longer than the other monomers, it is necessary to adjust them to a slightly higher concentration. For example, we adjust monomers 1, 6, 7, 12, 13, and 18 to 18 ng/ul and the other monomers to 15 ng/ul.

◆ CRITICAL STEP For subsequent digestion and ligation reactions, it is important that all monomers are at equimolar concentrations.

PAUSE POINT Amplified monomers can be stored at -20 ^{o}C for several months and can be re-used for assembling additional TALEs.

3.3.2 Construction of custom 20bp-targeting TALEs

TIMING 1.5 days (5 hr hands-on time)

10| *Select target sequence(s).* Typical TALE recognition sequences are identified in the 5' to 3' direction and begin with a 5' thymine. The procedure below describes the construction of TALEs that bind a 20 bp target sequence (5`-$T_0N_1N_2N_3N_4N_5N_6N_7N_8N_9N_{10}N_{11}N_{12}N_{13}N_{14}N_{15}N_{16}N_{17}N_{18}N_{19}$-3`, where N = A, G, T, or C), where the first base (typically a thymine) and the last base are specified by sequences

within the TALE backbone vector. The middle 18 bp are specified by the RVDs within the middle tandem repeat of 18 monomers according to the cipher NI = A, HD = C, NG = T, and NN = G or A. For targeting shorter or longer sequences, see Box 1.

11| *Divide target sequences into hexamers.* Divide N_1-N_{18} into sub-sequences of length 6 ($N_1N_2N_3N_4N_5N_6$, $N_7N_8N_9N_{10}N_{11}N_{12}$, and $N_{13}N_{14}N_{15}N_{16}N_{17}N_{18}$). For example, a TALE targeting 5`-TGAAGCACTTACTTTAGAAA-3` can be divided into hexamers as (T) GAAGCA CTTACT TTAGAA (A), where the initial thymine and final adenine (in parenthesis) are encoded by the appropriate backbone. In this example, the three hexamers will be: hexamer 1 = NN-NI-NI-NN-HD-NI, hexamer 2 = HD-NG-NG-NI-HD-NG, hexamer 3= NG-NG-NI-NN-NI-NI. Due to the adenine in the final position, we will use one of the NI backbones: pTALE-TF_v2(NI) or pTALEN_v2(NI).

12| *Assembling hexamers using Golden Gate digestion-ligation* (**Fig. 3.3**, stage 2). Prepare one reaction tube for each hexamer. Using the monomer plate schematic (**Fig. 3.4**), pipette 1ul of each normalized monomer into the corresponding hexamer reaction tube. Repeat this for all hexamers. For example, for the target from Step 10, set up tube 1 (1ul from each of G1, A2, A3, G4, E5, and A6), tube 2 (1 ul from each of E7, C8, C9, A10, E11, C12), and tube 3 (1 ul from each of D1, D2, B3, H4, B5, B6). To construct a TALE with 18 full repeats, 3 separate hexamer tubes are used.

◆ CRITICAL STEP Pay close attention when pipetting the monomers; it is very easy to accidentally pipette from the wrong well during this step.

13| To perform a simultaneous digestion-ligation (Golden Gate) reaction to assemble each hexamer (**Fig. 3.3**, stage 2) add the following reagents to each hexamer tube:

Component	**Amount**	**Final concentration**
Esp3I (BsmBI) 5 U/ul	0.75 ul	0.375 U/ul
Tango Buffer 10X	1 ul	1x
Dithiothreitol (DTT) 10mM	1ul	1 mM
T7 Ligase 3000 U/ul	0.25 ul	75 U/ul
ATP 10mM	1 ul	1 mM
	4 ul	
6 monomers	6 x 1 ul	
Total	10 ul	

◆ CRITICAL STEP Dithiothreitol (DTT) is easily oxidized in air. It should be freshly made or thawed from aliquots stored at -70 ^{o}C and used immediately.

14| Place each hexamer tube in a thermocycler to carry out the Golden Gate reactions using the following cycling conditions for ~ 3 hours:

Cycle number	**Digest**	**Ligate**
1-15	37°C, 5 min	20°C, 5 min
Hold at 4°C.		

PAUSE POINT This reaction can be left to run overnight.

15| Run out the ligation product on a gel to check for ~700 bp bands corresponding to the hexamer products (**Fig. 3.5a**, lane 7). Cast a 2% agarose gel in 1x TBE electrophoresis buffer with 2x SYBR Safe dye. The additional dye helps to visualize faint bands. The gel should have enough lanes to run out each Golden Gate reaction from Step 14; load 3 ul of each ligation product in separate lanes. Include in one lane 1 ug of the 1kb Plus DNA ladder. Run the gel at 15 V/cm until there is separation of the 650 bp ladder band from neighboring bands.

? TROUBLESHOOTING

16| *Exonuclease treatment to degrade non-circular ligation products* (**Fig. 3.3**, stage 3). During the Golden Gate reaction, only fully-ligated hexamers should be able to circularize. PlasmidSafe exonuclease selectively degrades non-circular (incomplete) ligation products. Add the following reagents to each hexamer reaction tube:

Component	**Amount**	**Final concentration**
PlasmidSafe DNAse 10U/ul	1 ul	0.66 U/ul
Plasmid-Safe 10X Reaction Buffer	1 ul	1x
ATP 10mM	1 ul	1 mM
	3 ul	
Golden gate reaction from Step 14	7 ul	
Total	10 ul	

17| Incubate each hexamer reaction tube with PlasmidSafe at 37°C for 30 minutes followed by inactivation at 70°C for 30 minutes.

PAUSE POINT After completion, the reaction can be frozen and continued later. The circular DNA should be stable for at least a week.

18| *Hexamer PCR* (**Fig. 3.3**, stage 4). Amplify each PlasmidSafe-treated hexamer in a 50 ul PCR reaction using high-fidelity Herculase II polymerase and the hexamer forward and reverse primers (Hex-F and Hex-R; Table 3.3). Add the following reagents to each PCR reaction:

Component	**Amount**	**Final concentration**
100mM dNTP (25mM each)	0.5 ul	1 mM
5X Herculase II reaction buffer	10 ul	1x
10uM each Hex-F and Hex-R primers	1 ul	200 nM
Herculase II Fusion DNA polymerase	0.5 ul	1x
Distilled water	37 ul	
	49ul	
PlasmidSafe-treated hexamer from Step 17	1 ul	
Total	50 ul	

19| Perform PCR on the reactions in Step 18 using the following cycling conditions:

Cycle number	**Denature**	**Anneal**	**Extend**
1	95°C, 2		

		min		
2-36		95°C, 20 s	60°C, 20 s	72°C, 30 s
37				72°C, 3 min

20| *Gel purification of amplified hexamers.* Due to the highly repetitive template, it is necessary to purify the amplified hexamer product from the other amplicons. Cast a 2% agarose gel in 1x TBE electrophoresis buffer with 1x SYBR Safe dye. The gel should have enough lanes to run out each PCR product from Step 19 and the comb size should be big enough to load 40-50ul of PCR product. Include in one lane 1 ug of the 1kb Plus DNA ladder. Run the gel at 15 V/cm until there is separation of the 650 bp ladder band from neighboring bands. Using a clean razor blade, excise each hexamer band, which should be nearly aligned with the 650bp band from the ladder (**Fig. 3.5**, lane 9).

◆ CRITICAL STEP Avoid any cross-contamination by ethanol sterilization of work surfaces, razor blades, etc. during the gel extraction and between each individual band excision.

! CAUTION Wear appropriate personal protective equipment, including a facemask, when performing gel stabs to minimize risks associated with prolonged light or mutagenic DNA dye exposure.

? TROUBLESHOOTING

21| Purify the hexamer gel bands from Step 20 using the MinElute Gel Extraction kit following the manufacturer's directions. Elute the DNA from each reaction using 20 ul of Buffer EB prewarmed to 55°C.

22| *Gel normalization of purified hexamer concentrations.* Cast a 2% agarose gel in 1x TBE electrophoresis buffer with 1x SYBR Safe dye. The gel should have enough lanes to run out 2 ul of each purified hexamer from Step 21. Include in one lane 10 ul of the quantitative DNA ladder. Run the gel at 15V/cm until all lanes of the quantitative ladder are clearly separated. Each hexamer lane should contain only a single (purified) band.

23| Image the gel using a quantitative gel imaging system. Each lane should have only the ~700 bp hexamer product. Make sure the exposure is short enough so that none of the bands are saturated.

24| Quantify the integrated intensity of each hexamer band using ImageJ or other gel quantification software. Use the quantitative ladder with known concentrations (5, 10, 20, 40, 100 ng) to generate a linear fit and quantify the concentration of each purified hexamer.

? TROUBLESHOOTING

25| Adjust the concentration of each hexamer to 20 ng/ul by adding Buffer EB.

26| *Golden Gate assembly of hexamers into TALE backbone* (**Fig. 3.3**, stage 5). Combine the hexamers and the appropriate TALE backbone vector (transcription factor or nuclease) in a Golden Gate digestion-ligation. For example, we will use a TALE backbone with NI as the 0.5 repeat for the target sequence in Step 10 since N_{19}=A. For this ligation, a 1:1 molar ratio of insert:vector works well. Set up one reaction tube for each TALE. Also,

prepare a negative control ligation by including the TALE backbone vector without any hexamers.

Component	TALE	Negative control	Final concentration
TALE backbone vector (100ng/ul)	1 ul	1 ul	10 ng/ul
BsaI-HF (20 U/ul)	0.75 ul	0.75 ul	1.5 U/ul
10x NEBuffer 4	1 ul	1 ul	1x
10x Bovine serum albumin	1 ul	1 ul	1x
ATP 10mM	1 ul	1 ul	1 mM
T7 Ligase (3000 U/ul)	0.25 ul	0.25 ul	75 U/ul
	5 ul	5 ul	
3 purified hexamers (20 ng/ul)	3 ul (1 ul each)		2 ng/ul each
Distilled water	2 ul	5 ul	
Total	10 ul	10 ul	

◆ CRITICAL STEP As a negative control, set up a separate reaction omitting the purified hexamers (i.e. including only the TALEN or TALE-TF backbone).

27| Place the tubes from Step 26 in a thermocycler to carry out the Golden Gate reactions using the following cycling conditions for ~4 hours:

Cycle number	Digest	Ligate	Inactivate
1-20	37°C, 5 min	20°C, 5 min	
21			80°C, 20 min

PAUSE POINT Ligation products can be frozen at -20 °C and stored at least one month for transformation into bacteria at a later time.

28| Although it is not necessary, it is possible to run out the ligation product on a gel to check for ~1.8 kbp band corresponding to the properly assembled 18mer tandem repeat. To check the ligation product, cast a 2% agarose gel in 1x TBE electrophoresis buffer with 2x SYBR Safe dye. The additional dye helps to visualize faint bands. Load 5ul of the ligation product from Step 27. Include in one lane 1 ug of the 1kb Plus DNA ladder. Run the gel at 15 V/cm until there is clear separation of the 1650 bp and 2000 bp ladder bands. Alternatively, proceed directly to transformation (Step 29) without running a gel; transformation is very sensitive and, even when a clear band cannot be visualized on the gel, there is often enough plasmid for transformation of high competency cells.

? TROUBLESHOOTING

3.3.3 Verifying correct TALE repeat assembly

TIMING 3 days (4 hr hands-on time)

29| *Transformation.* Transform the ligation products from Step 27 into a competent *E. coli*; in our lab, we use Stbl3 for routine transformation. Transformation can be done following the protocol supplied with the cells. Briefly, add 5 ul of the ligation product to 50 ul of ice-cold chemically competent Stbl3 cells, incubate on ice for 5 min, incubate at 42°C for 45 sec, return immediately to ice for 5 min, add 250 ul of SOC medium, incubate at 37°C for 1 hr on a shaking incubator (250 rpm), plate 100 ul of transformation on a LB plate containing 100 ug ml^{-1} ampicillin and incubate overnight at 37°C.

30| Inspect all plates from Step 29 for bacterial colony growth. Typically, we see few colonies on the negative control plates (only backbone in the Golden Gate digestion-ligation) and tens to hundreds of colonies on the complete TALE ligation plates.

? TROUBLESHOOTING

31| For each TALE plate, pick 8 colonies to check the assembly fidelity. Using a sterile 20 ul pipette tip, touch the tip to a single colony, streak onto a single square on a pre-warmed, new gridded LB-ampicillin plate to save the colony, and then swirl the tip in 100ul of distilled water to dissolve the colony for colony PCR. Repeat this procedure for all colonies to be checked, streaking each new colony into a separate square on the gridded LB-ampicillin plate. After finishing, incubate the gridded plate at 37°C for at least 4 hours to grow up the colony streaks.

32| *Colony PCR.* Using the colonies selected in Step 31 as templates, set up colony PCR to verify that the correctly assembled tandem 18mer repeat has been ligated into the TALE backbone. We have found that the colony PCR reaction is sensitive to excessive template concentration; therefore we typically use 1 ul of the 100 ul colony suspension from Step 31. For colony PCR, use primers TALE-Seq-F1 and TALE-Seq-R1 for amplification (**Table 2**). Set up the following colony PCR reaction:

Component	**Amount**	**Final concentration**
Colony suspension from Step 31	1 ul	

100mM dNTP (25mM each)	0.25 ul	1 mM
10x *Taq-B* polymerase buffer	2.5 ul	1x
10uM each TALE-Seq-F1 and TALE-Seq-R1 primers	0.25 ul	100 nM
Taq-B polymerase (5 U/ul)	0.1 ul	0.02 U/ul
Distilled water	20.9 ul	
Total	25 ul	

33| Perform colony PCR on the reactions in Step 32 using the following cycling conditions:

Cycle number	**Denature**	**Anneal**	**Extend**
1	94°C, 3 min		
2-31	94°C, 30 s	60°C, 30 s	68°C, 2 min
32			68°C, 5 min

34| To check the colony PCR result, cast a 1% agarose gel in 1x TBE electrophoresis buffer with 1x SYBR Safe dye. The gel should have enough lanes to run out 10 ul of each PCR product from Step 33. Include in one lane 1 ug of the 1 kb Plus DNA Ladder. Run the gel at 15 V/cm until there is clear separation of the 1650 bp and 2000 bp ladder bands.

35| Image the gel and identify which colonies have the correct insert size. For an insert of 18 monomers (3 hexamers ligated into the TALE backbone vector), the product should be a single band of size 2175 bp (**Fig. 3.5b**, lane 1). Incorrect ligation products will show

bands of different sizes. In place of colony PCR, plasmid DNA from prepared clones can be digested with *Afe*I. In both backbones (TALE-TF and TALEN), *Afe*I cuts 4 times. For both backbones, one fragment contains the entire tandem repeat region and should be of size 2118 bp for a correctly assembled 18mer. For the TALE-TF backbone, the correct clone will produce 4 bands with sizes: 165bp, 2118bp, 3435bp and 3544bp (**Fig. 3.5b**, lane 2). The 3435bp and 3544bp bands are difficult to separate on a 1% agarose gel and therefore a correct clone will show three bands with the middle 2118bp band indicating an intact tandem 18mer repeat (**Fig. 3.5b**, lane 2). For the TALEN backbone, the correct clone will produce 4 bands with sizes: 165 bp, 2118 bp, 2803 bp, 3236 bp.

? TROUBLESHOOTING

36| *Miniprep and sequencing.* For each clone with the correct band size, inoculate a colony from the gridded plate into 3 ml of LB media with 100 ug ml^{-1} ampicillin and incubate at 37°C in a shaking incubator overnight.

37| Isolate plasmid DNA from overnight cultures using a QIAprep Spin Miniprep Kit following the manufacturer's instructions.

38| Verify the sequence of each clone by sequencing the tandem repeat region using sequencing primers (see **Table 3.3**) TALE-Seq-F1 (forward primer annealing just before the first monomer), TALE-Seq-F2 (forward primer annealing at the beginning of the seventh monomer) and TALE-Seq-R1 (reverse primer annealing after the final 0.5 monomer). For most TALEs, reads from all 3 primers are necessary to unambiguously

verify the entire sequence. Verify the sequencing result using our online, freely-available TALE software (http://taleffectors.com/tools/) or using standard sequence alignment methods (e.g. ClustalW). After entering the target site sequence, our software generates a TALE-TF or TALEN reference sequence in either FASTA format or as an annotated GenBank vector map (*.gb file) that can be viewed using standard plasmid editor software (e.g. everyVECTOR, VectorNTI, or LaserGene SeqBuilder). The software also aligns sequencing reads (entered in FASTA format) to the generated reference sequence to allow for easy clone verification. Detailed instructions can be found on our website.

? TROUBLESHOOTING

3.3.4 Transfection of TALE-TF and TALEN into HEK293FT cells

TIMING 2 days (1 hour hands-on time)

39| Plate HEK293FT cells onto 6-well plates in D10 culture medium without antibiotics approximately 24h prior to transfection at a seeding density of around 1×10^6 cells per well and a seeding volume of 2mL. Scale up and down the culture according to the manufacturer's manual provided with the 293FT cells if needed.

40| *Prepare DNA for transfection.* Quantify the DNA concentration of the TALE plasmids used for transfection using reliable methods (such as UV spectrophotometry or gel quantification).

◆ CRITICAL STEP: The DNA concentration of the TALE plasmids should be quantified to guarantee that an accurate amount of TALE DNA will be used during the transfection.

41| Prepare the DNA-Opti-MEM mix as follows using option A if testing transcriptional modulation, or option B if testing nuclease activity:

A. DNA-Opti-MEM mix for testing transcriptional modulation.

i) Mix 4 μg of TALE-TF plasmid DNA with 250 μl of Opti-MEM medium. Include controls (e.g. RFP plasmid or mock transfection) to monitor transfection efficiency and cell health respectively.

B. DNA-Opti-MEM mix for testing nuclease activity.

i) Mix 2 μg of the Left and 2 μg of the Right TALEN (**Figure 3.1d**) plasmid DNA with 250 μl of Opti-MEM medium. Control transfections should be done by omitting one or both of the TALENs. Also include controls (e.g. an RFP plasmid or mock transfection) to monitor transfection efficiency and cell health respectively. For all transfections, make sure the total amount of DNA transfected is the same across conditions – when omitting one or both TALENs, supplement with empty vector DNA to maintain the same total DNA amount.

42| Prepare the Lipofectamine-Opti-MEM solution by diluting 10 μl of Lipofectamine 2000 with 250 μl of Opti-MEM. Mix the solution thoroughly by tapping the tube and incubating for 5 minutes at room temperature.

43| Add the Lipofectamine-Opti-MEM solution to the DNA-Opti-MEM solution to form the DNA-Lipofectamine complex. Mix well by gently pipetting up and down. Incubate for 20 minutes at room temperature.

◆ CRITICAL STEP Make sure the complex is thoroughly mixed. Insufficient mixing results in lower transfection efficiency.

◆ PAUSE POINT The transfection complex will remain stable for 6 hours at room temperature.

44| Add 500 μl of the DNA-Lipofectamine complex to each well of the 6-well plate from Step 39 directly. Mix gently by rocking the plates back and forth.

45| Incubate cells at 37°C with 5% CO_2 for 24 hours. At this point, determine the transfection efficiency by estimating the fraction of fluorescent cells in the positive control transfection (e.g. RFP plasmid) using a fluorescence microscope.

CRITICAL STEP If incubation beyond 48 hours is needed, change the culture medium with fresh D10 supplemented with antibiotics on a daily basis. This will not affect the transfection efficiency.

? TROUBLESHOOTING

3.3.5 TALE functional characterization

46| To measure TALEN cutting efficiency using Surveyor nuclease follow option A, or to measure TALE-TF transcriptional activation using qRT-PCR follow option B:

A. Measuring TALEN cutting efficiency using Surveyor nuclease

TIMING 6 hr (3 hr hands-on time)

i) Remove culture medium from each well from Step 45 and add 100 μl of QuickExtract DNA Extraction Solution to each well and pipette thoroughly to lyse cells. Transfer the lysate to a PCR tube.

ii) Extract DNA from the lysate from Step 46Ai using the following cycling conditions:

Cycle number	**Condition**
1	68°C, 15 min
2	95°C, 8 min

iii) *PCR amplification of region surrounding TALEN target site.* Prepare the following PCR reaction using the genomic DNA from Step 46Aii:

Component	**Amount**	**Final concentration**
gDNA from Step 46Aii	0.5 ul	
100mM dNTP (25mM each)	0.5 ul	1 mM
5X Herculase II reaction buffer	10 ul	1x
10uM each of target-specific Surveyor forward and reverse primers (see Experimental Design)	1 ul	200 nM
Herculase II Fusion DNA polymerase	0.5 ul	1x
Distilled water	37.5 ul	
Total	50 ul	

◆ CRITICAL STEP Surveyor procedure (Steps 46Aiii-xv) is carried out according to the manufacturer's protocol and is described in greater detail in the Surveyor manual. We provide brief details here since mutation detection by mismatch endonuclease is not a common procedure for most laboratories.

◆ CRITICAL STEP When performing the Surveyor assay for the first time, we suggest carrying out the positive control reaction included with the Surveyor nuclease kit.

iv) Perform PCR using the following cycling conditions:

Cycle number	**Denature**	**Anneal**	**Extend**
1	95°C, 3 min		
2-36	95°C, 30 s	55°C, 15 s	72°C, 30s
37			72°C, 5 min

v) Check the PCR result by running 5 ul of PCR product on a 2% agarose gel in 1x TBE electrophoresis buffer with 1x SYBR Safe dye. Include in one lane 10 ul of the quantitative DNA ladder. Run the gel at 15 V/cm until all bands are clearly separated. For all templates, it is important to make sure that there is only a single band corresponding to the intended product for the primer pair. The size of this band should be the same as calculated from the distance between the two primer annealing sites in the genome.

◆ CRITICAL STEP If multiple amplicons are generated from the PCR reaction, re-design primers and re-optimize the PCR conditions to avoid off-target amplification.

? TROUBLESHOOTING

vi) Image the gel using a quantitative gel imaging system. Make sure the exposure is short enough so that none of the bands are saturated. Quantify the integrated intensity of each PCR product using ImageJ or other gel quantification software. Use the quantitative ladder with

known concentrations (5, 10, 20, 40, 100 ng) to generate a linear fit. Adjust the DNA concentration of the PCR product by diluting with 1x Herculase II reaction buffer so that it is in the range of 25 - 80 ng/µl.

vii) *DNA heteroduplex formation.* At this point, the amplified PCR product includes a mixture of both modified and unmodified genomic DNA (TALEN-modified DNA will have a few bases of sequence deletion near the TALEN cut site due to exonuclease activity during NHEJ). For Surveyor mismatch detection, this mixture of products must first be melted and re-annealed such that heteroduplexes are formed. DNA heteroduplexes contain strands of DNA that are slightly different but annealed (imperfectly) together. Given the presence of both unmodified and modified DNA in a sample, a heteroduplex may include one strand of unmodified DNA and one strand of TALEN-modified DNA. Heteroduplexes can also be formed from re-annealing of two different TALEN-modified products as NHEJ exonuclease activity can produce different mutations. To cross-hybridize wild type and TALEN-modified PCR products into hetero- and homoduplexes, all strands are melted and then slowly re-annealed (**Figure 3.6a**). Place 300 ng of the PCR product from Step 46Avi in a thermocycler tube and bring to a total volume of 20 µl with 1x Herculase II reaction buffer.

viii) Perform cross-hybridization on the diluted PCR amplicon from Step 46Avii using the following cycling conditions:

Cycle number	**Condition**
1	95°C, 10 min
2	95°C to 85°C, -2°C/s

3	85°C, 1 min
4	85°C to 75°C, -0.3°C/s
5	75°C, 1 min
6	75°C to 65°C, -0.3°C/s
7	65°C, 1 min
8	65°C to 55°C, -0.3°C/s
9	55°C, 1 min
10	55°C to 45°C, -0.3°C/s
11	45°C, 1 min
12	45°C to 35°C, -0.3°C/s
13	35°C, 1 min
14	35°C to 25°C, -0.3°C/s
15	25°C, 1 min

ix) *Surveyor Nuclease S digestion.* To treat the cross-hybridized homo- and hetero-duplexes using Surveyor Nuclease S to determine TALEN cleavage efficiency (**Figure 3.6a**), add the following components together on ice and mix by pipetting gently:

Component	**Amount**	**Final concentration**
0.15 M $MgCl_2$ solution	2 ul	15 mM
Surveyor Nuclease S	1 ul	1x
Surveyor Enhancer S	1 ul	1x
	4 ul	

Re-annealed duplexes from Step 46Aviii	16 ul
Total	20 ul

x) Incubate the reaction from Step 46Aix at 42°C for 1 hour.

xi) Add 2 ul of the Stop Solution from the Surveyor kit.

◆ PAUSE POINT: The digestion product can be stored at -20°C for analysis at a later time.

xii) Cast a 2% agarose gel in 1x TBE electrophoresis buffer with 1x SYBR Safe dye. When casting the gel, it is preferable to use a thin comb size (<1 mm) for the sharpest possible bands. The gel should have enough lanes to run out 20 ul of each digestion product band from Step 46Axi. Include in one lane 1 ug of the 1kb Plus DNA ladder. Run the gel at 5 V/cm until the Orange G loading dye has migrated 2/3rds of the way down the gel.

xiii) Image the gel using a quantitative gel imaging system. Make sure the exposure is short enough so that none of the bands are saturated. Each lane from samples transfected with both left and right TALENs should have a larger band corresponding to the uncut genomic amplicon (the same size as in Step46Av) and smaller bands corresponding to the DNA fragments resulting from the cleavage of the genomic amplicon by Surveyor nuclease. Controls (no transfection, control plasmid transfection, or transfection omitting one of the TALENs) should only have the larger band corresponding to the uncut genomic amplicon.

? TROUBLESHOOTING

xiv) Quantify the integrated intensity of each band using ImageJ or other gel quantification software. For each lane, calculate the fraction of the PCR product cleaved (f_{cut}) using the following formula: $f_{cut} = a / (a+b)$, where a = the integrated intensity of both of the cleavage product bands, and b = the integrated intensity of uncleaved PCR product band. A sample Surveyor gel for TALENs targeting human AAVS1 is shown in **Figure 3.6b**.

xv) Estimate the percentage of TALEN-mediated gene modification using the following formula(*93*):

$$100 \times (1-(1-f_{cut})^{1/2})$$

This calculation can be derived from the binomial probability distribution given a few conditions: that strand reassortment during the duplex formation is random, that there is a negligible probability of the identical mutations reannealing during duplex formation, and that the Surveyor nuclease digestion is complete.

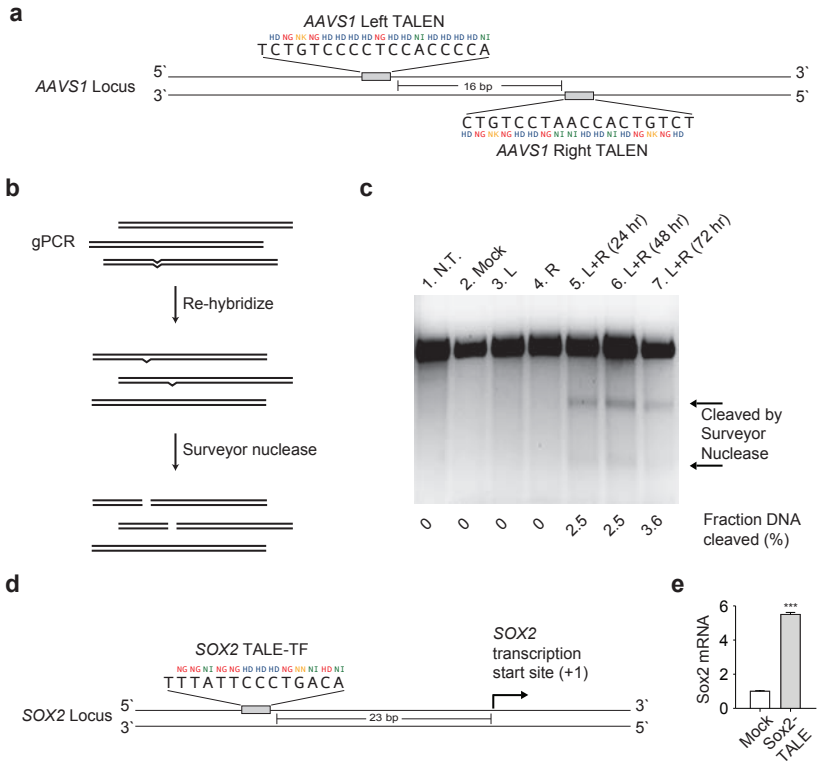


Figure 3.6 Examples of TALE-TF and TALEN activity in 293FT cells. (a) Schematic of the Surveyor nuclease assay used to determine TALEN cleavage efficiency. First, genomic PCR is used to amplify the TALEN target region from a heterogeneous population of TALEN-modified and unmodified cells, and the gPCR products are re-annealed slowly to generate heteroduplexes. The re-annealed heteroduplexes are cleaved by Surveyor nuclease while homoduplexes are left intact. TALEN cleavage efficiency is calculated based on the fraction of cleaved DNA. (**b**) Gel showing the Surveyor nuclease result from the AAVS1 TALEN pair (from **Fig. 3.1d**). Lanes 1—4: controls from un-transfected (N.T.) cells and cells transfected with a plasmid carrying GFP (Mock), AAVS1 left TALEN only (L), and AAVS1 right TALEN only (R). Lanes 5—7: cells transfected with AAVS1 Left and Right TALENs (L+R) for 24, 48, and 72 hours. The two lower bands indicated by the arrows are Surveyor-cleaved DNA products. (**c**) 293FT cells transfected with the *SOX2* TALE-TF (from **Fig. 3.1c**) exhibited a 5 fold increase in the amount of *SOX2* mRNA compared with mock transfected cells. Error bars indicate s.e.m.; n = 3. *** indicates $P < 0.005$. Panel c was modified with permission from *Nature Biotechnology (27)* (*Nature Biotechnology* (c) 2011, Macmillian Publishers Ltd.).

B. Measuring TALE-TF transcriptional activation using qRT-PCR

TIMING 5 hr (3 hr hands-on time)

i) *RNA extraction.* Aspirate the medium in each well of the 6-well plates from Step 45 at 72 hours after transfection.

◆ CRITICAL STEP Use proper RNA handling techniques to prevent RNA degradation, including cleaning bench surfaces and pipettes with RNAseZAP. Use RNAse-free consumables and reagents.

ii) Wash the cells in each well twice with 1 ml of DPBS.

iii) Harvest ~1×10^6 cells for subsequent total RNA extraction by trypsinizing the cells with 500 μl trypsin with EDTA. Incubate for 1-2 minutes to let the cells detach from the bottom of the wells.

◆ CRITICAL STEP Do not leave the cells in trypsin for longer than a few minutes.

iv) Neutralize the trypsin by adding 2ml of D10 medium.

v) In a 15ml centrifuge tube, centrifuge the cell suspension at 300×g for 5 min. Carefully aspirate all of the supernatant.

◆ CRITICAL STEP Incomplete removal of the supernatant can result in inhibition of cell lysis.

PAUSE POINT: Cells can be frozen at -80°C for 24 hours.

vi) Extract and purify RNA from the cells in Step 46Bv using the RNeasy Mini Kit and QIAshredder following the manufacturer's directions. Elute the RNA from each column using 30 ul of nuclease-free water.

vii) Measure the RNA concentration using a UV spectrophotometer.

viii) *cDNA reverse-transcription.* Generate cDNA using the iScript cDNA Synthesis Kit following the manufacturer's directions. For matched negative controls, perform the reverse transcription without the reverse-transcriptase enzyme.

ix) *Quantitative PCR.* Thaw on ice the appropriate TaqMan probe for the target gene and for an endogenous control gene.

◆ CRITICAL STEP Protect the probes from light and do not allow the thawed probes to stay on ice for an extended time.

x) Following the TaqMan Universal PCR Master Mix manufacturer's directions, prepare 4 technical replicate qPCR reactions for each sample in optical thermocycler strip tubes or 96-well plates. Set up negative controls for non-specific amplification as indicated in the directions: namely, RNA template processed without reverse transcriptase ("no RT") and a no-template control.

xi) Briefly centrifuge the samples to remove any bubbles and amplify them in a TaqMan-compatible qRT-PCR machine with the following cycling parameters.

Cycle number	Denature	Anneal and Extend
1	95°C, 20 s	
2-41	95°C, 1 s	60°C, 20 s

xii) Analyze data and calculate the level of gene activation using the $\Delta\Delta C_T$ method(*92, 101*). TALE-TF results from qRT-PCR assay of *SOX2* activation in HEK293 cells are shown in **Figure 3.6c**.

◆ CRITICAL STEP The $\Delta\Delta C_T$ method assumes that amplification efficiency is 100% (ie. number of amplicons doubles after each cycle). For new probes (such as custom TaqMan probes), amplification from a template dilution series (spanning at least 5 orders of magnitude) should be performed to characterize amplification efficiency. For standard TaqMan Gene Expression Assay probes, this is not necessary as they are designed to have 100±10% amplification efficiency.

? TROUBLESHOOTING

• TIMING

Steps 1-9, Monomer library amplification and normalization: 6 hr

Steps 10-28, TALE hierarchical ligation assembly: 1.5 days (5 hr hands-on time)

Steps 29-38, TALE transformation and sequence verification: 3 days (4 hr hands-on time)

Steps 39-45, Transfection of TALE-TF and TALEN into HEK293FT cells: 2 days (1 hr hands-on time)

Steps 46A and B, TALE functional characterization with RT-qPCR or Surveyor: 6 hr (3 hr hands-on time)

3.4 Troubleshooting

Troubleshooting advice can be found in **Table 3.2.**

Table 3.2 Troubleshooting

Step	Problem	Possible reason	Solution
4	Uneven amplification across monomers	Not using Herculase 2 Fusion polymerase	Optimize annealing temperature and Mg^{2+} and DMSO concentrations
8	Low DNA concentration after elution	Residual ethanol on purification column	Air dry columns before elution at 37°C for a longer period of time
		Incorrect vacuum pressure during DNA binding	Adjust vacuum pressure according to the manufacturer's suggestions
15	No visible hexamer band (~700 bp)	Not adding equimolar amounts of monomers	Gel normalize monomer concentration
		Degraded DTT or ATP	Use fresh stocks of DTT and ATP, which degrade easily

(Table 3.2, continued)

	No visible hexamer band (~700 bp) but smaller bands present	Wrong monomer(s) added during pipetting	Re-select monomers
		Monomer concentration is too low	Increase the number of Golden Gate digestion-ligation cycles and/or increase the concentration of monomers to >20 ng/ul; there is no detrimental effect to using more monomers in an equimolar ratio
20	No visible hexamer band (~700 bp)	Unsuccessful Golden Gate digestion-ligation	Verify on a gel that the Golden Gate digestion-ligation product from Step 15 is visible; increase monomer concentration
24	Low concentration for purified hexamers	Unsuccessful gel extraction	Ensure that there is no residual ethanol during elution or increase PCR reaction volume
28	No visible 18mer band (~1.8 kbp)	Unsuccessful Golden Gate digestion-ligation	Increase hexamer concentration in Golden Gate digestion-ligation in Step 26 or proceed directly to transformation in Step 29

(Table 3.2, continued)

30	More than a few colonies on negative control plate	Compromised TALE backbone	Perform a restriction digest of the backbone to verify integrity
35	Colony PCR bands are smeared	Too much template	Dilute colony suspension 10x-100x
38	Monomers assembled in incorrect order	Misligation	Misligation occurs at a very low frequency; analyze two additional clones
45	Low transfection efficiency	Low DNA quality	Prepare DNA using high quality plasmid preparation
		Suboptimal DNA to Lipofectamine2000 ratio	Titrate DNA to Lipofectamine2000 ratio to determine optimal transfection condition
46Av	Multiple amplicons	Nonspecific primers	Design new primers and verify specificity using PrimerBLAST; use touchdown PCR
	No amplification	Suboptimal PCR condition	Optimize annealing temperature and Mg^{2+} and DMSO concentrations
46Axiii	No cleavage bands visible	TALEN unable to cleave target site	Design new TALEN pairs targeting nearby sequences

(Table 3.2, continued)

46Bxii	No increase in transcription in target mRNA	TALE-TF unable to access target site	Design new TALE-TFs targeting nearby sequences

3.5 Anticipated results

TALE-TFs and TALENs can facilitate site-specific transcriptional modulation(*27, 53, 54, 57*) and genome editing(*39, 53, 56, 58, 60-63*) (**Table 3.1**). TALENs can be readily designed to introduce double-stranded breaks at specific genomic loci with high efficiency. In our experience, a pair of TALENs designed to target the human AAVS1 locus is able to achieve up to 3.6% cutting efficiency in 293FT cells as determined by Surveyor nuclease assay (**Fig. 3.6a-b**). TALE-TFs can also robustly increase the mRNA levels of endogenous genes. For example, a TALE-TF designed to target the proximal promoter region of *SOX2* in human cells is able to elevate the level of endogenous *SOX2* gene expression by up to 5 fold(*27*) (**Fig. 3.6c**). The ability for TALE-TFs and TALENs to act at endogenous genomic loci is dependent on the chromatin state as well as yet-to-be-determined mechanisms regulating TALE DNA binding(*102, 103*). For these reasons we typically build several TALE-TFs or TALEN pairs for each genomic locus we aim to target. These TALE-TFs and TALENs are designed to bind neighboring regions around a specific target site since some binding sites might be more accessible than others. The reason why some TALEs exhibit significantly lower levels of activity remains unknown, though it is likely due to position- or cell state-specific epigenetic modifications preventing access to the binding site. Due to differences in epigenetic states

between different cells, it is possible that TALEs that fail to work in a particular cell type might work in a different cell type.

Table 3.3 Primer sequences for TALE construction

Name	Sequence	Purpose
Ex-F1	TGCGTCcgtctcCGAACCTTAAACCGGCCAACATACCggtctcCTGACCCCAGAGCAGGTCGTG	monomer amplification
Ex-F2	TGCGTCcgtctcCGAACCTTAAACCGGCCAACATACCggtctcGACTTACACCCGAACAAGTCGTGGCAATTGCGAGC	
Ex-F3	TGCGTCcgtctcCGAACCTTAAACCGGCCAACATACCggtctcGCGGCCTCACCCCAGAGCAGGTCG	
Ex-F4	TGCGTCcgtctcCGAACCTTAAACCGGCCAACATACCggtctcGTGGGCTCACCCCAGAGCAGGTCG	
Ex-R1	GCTGACcgtctcCGTTCAGTCTGTCTTTCCCCTTTCCggtctcTAAGTCCGTGCGCTTGGCAC	
Ex-R2	GCTGACcgtctcCGTTCAGTCTGTCTTTCCCCTTTCCggtctcAGCCGTGCGCTTGGCACAG	
Ex-R3	GCTGACcgtctcCGTTCAGTCTGTCTTTCCCCTTTCCggtctcTCCCATGGGCCTGACATAACACAGGCAGCAACCTCTG	
Ex-R4	GCTGACcgtctcCGTTCAGTCTGTCTTTCCCCTTTCCggtctcTTAGACCGTGCGCTTGGCACAG	
In-F2	CTTGTTATGGACGAGTTGCCcgtctcGTACGCCAGAGCAGGTCGTGGC	
In-F3	CCAAAGATTCAACCGTCCTGcgtctcGAACCCCAGAGCAGGTCGTG	
In-F4	TATTCATGCTTGGACGGACTcgtctcGGTTGACCCCAGAGCAGGTCGTG	
In-F5	GTCCTAGTGAGGAATACCGGcgtctcGCCTGACCCCAGAGCAGGTCGTG	
In-F6	TTCCTTGATACCGTAGCTCGcgtctcGGACACCAGAGCAGGTCGTGGC	
In-R1	TCTTATCGGTGCTTCGTTCTcgtctcCCGTAAGTCCGTGCGCTTGGCAC	
In-R2	CGTTTCTTTCCGGTCGTTAGcgtctcTGGTTAGTCCGTGCGCTTGGCAC	
In-R3	TGAGCCTTATGATTTCCCGTcgtctcTCAACCCGTGCGCTTGGCACAG	
In-R4	AGTCTGTCTTTCCCCTTTCCcgtctcTCAGGCCGTGCGCTTGGCACAG	
In-R5	CCGAAGAATCGCAGATCCTAcgtctcTTGTCAGTCCGTGCGCTTGGCAC	
Hex-F	CTTAAACCGGCCAACATACC	hexamer amplification
Hex-R	AGTCTGTCTTTCCCCTTTCC	
TALE-Seq-F1 (aka colony PCR forward)	CCAGTTGCTGAAGATCGCGAAGC	sequencing forward primer used to check monomers 1-6.
TALE-Seq-F2	ACTTACACCCGAACAAGTCG	sequencing forward primer used to check monomers 7-12
TALE-Seq-R1 (aka colony PCR reverse)	TGCCACTCGATGTGATGTCCTC	sequencing primer used to check monomers 13-18 for TALEs with less than 18 full monomer repeats, and used to check monomers 19-24 for TALEs with more than 18 monomers.
TALE-Seq-R2	CCCATGGGCCTGACATAA	sequencing reverse primer used to check monomers 13-18 in TALEs with more than 18 full monomer repeats

Page intentionally left blank.

4. Comprehensive Interrogation of Natural TALE DNA Binding Modules and Transcriptional Repressor Domains

The majority of the work presented here is done by Le Cong, with help from Dr. Ruhong Zhou on the computational work, Yu-chi Kuo, and Margaret Cunniff on the molecular biology work. The contents in this chapter has been pubished as Cong L, et al. Nature Communications. 2012 (*29*).

4.1 Introduction

Transcription activator-like effectors (TALEs) are bacterial effector proteins found in *Xanthamonas sp.* and *Ralstonia sp*. Each TALE contains a DNA binding domain consisting of 34 amino acid tandem repeat modules, where the 12^{th} and 13^{th} residues of each module, referred to as repeat variable diresidues (RVDs), specify the target DNA base(*23, 25*). Four of the most abundant RVDs from naturally occurring TALEs have established a simple code for DNA recognition (e.g. NI for adenine, HD for cytosine, NG for thymine, and NN for guanine or adenine)(*23, 25*). Using this simple code, TALEs have been developed into a versatile platform for achieving precise genomic and transcriptomic perturbations across a diverse range of biological systems(*27, 28, 53, 54, 57, 104*). However, two limitations remain: first, there lacks a RVD capable of robustly and specifically recognizing the DNA base guanine, a highly prevalent base in mammalian genomes(*105*); and second, a viable TALE transcriptional repressor for mammalian applications has remained elusive, which is highly desirable for a variety of synthetic biology and disease-modeling applications(*105*). To address these two limitations, we conducted a series of screens and found that: first, of all naturally

occurring TALE RVDs, the previously unidentified RVD Asn-His (NH) can be used to achieve guanine-specific recognition; and second, the mSin Interaction Domain (SID)(*106*) can be fused to TALEs to facilitate targeted transcriptional repression of endogenous mammalian gene expression. These advances further improve the power and precision of TALE-based genome engineering technologies, enabling efficient bimodal control of mammalian transcriptional processes.

4.2 Results

4.2.1 Screening of novel TALE RVDs

Previously, the RVD NK was reported to have more specificity for guanine than NN(*54*). However, recent studies have show that substitution of NK with NN leads to substantially lower levels of activity(*87*). To identify a more specific guanine-binding RVD with higher biological activity, we identified and evaluated a total of 23 naturally occurring RVDs (Fig. 4.1) from the set of known *Xanthomonas* TALE sequences in Genbank. In order to directly compare the DNA binding specificity and activity of all RVDs in an unbiased manner, we designed a set of 2312.5-repeat TALEs where we systematically substituted RVDs 5 and 6 with the 23 naturally occurring RVDs (RVD-TALEs; Fig. 4.1a). This design allowed us to maintain a consistent RVD context surrounding the two varied RVD positions. Additionally, we fused a *Gaussian* luciferase gene (Gluc) with a 2A peptide linker to the RVD-TALEs to control for the differences in TALE protein expression levels (Fig. 4.1a). We used each RVD-TALE (e.g. NI-TALE, HD-TALE, etc.) to assess the base-preference and activity strength of its corresponding RVD – this is measured by comparing each RVD-TALE's ability to activate transcription from each of

the four base-specific *Cypridina* luciferase reporter (Cluc) plasmids with A, G, T, and C substituted in the 6th and 7th positions of the TALE binding site (A-, G-, T-, or C-reporters; Fig. 4.1a).

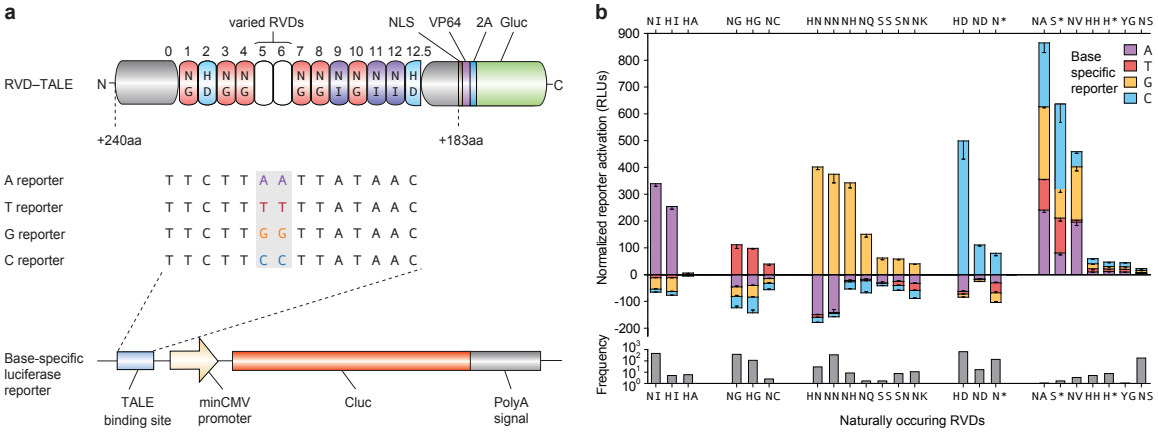


Figure 4.1 Identification of an optimal guanine-specific repeat variable diresidue (RVD). a, Design of the TALE RVD screening system. Each RVD screening TALE (RVD-TALE) contains 12.5 repeats with RVDs 5 and 6 substituted with the 23 naturally occurring RVDs, and is fused to a *Gaussia* luciferase gene via a 2A peptide linker. The truncations used for the TALE is marked at the N- and C- termini with numbers of amino acids retained (top). Four different base-specific reporters with A, T, G, and C substituted in the 6th and 7th nucleotides of the binding site are used to determine the base-specificity of each RVD (middle). Each reporter is constructed by placing the TALE binding site upstream of a minimal CMV promoter driving *Cypridina* luciferase (bottom). **b,** Base-preference of each natural RVD (top) is determined by measuring the levels of relative luminescence unit (RLU) for each base-specific reporter after background subtraction and normalization based on TALE protein expression level (top). We clustered RVDs according to their base-preference after performing one-way analysis of variance (ANOVA) tests on each RVD. For RVDs with a single statistically significant reporter activity ($p < 0.05$, one-way ANOVA), we plotted the reporter activity of the

(Figure 4.1, continued) preferred base above the x-axis, whereas the reporter activities for the non-preferred bases are shown below the x-axis as negative. We clustered and ranked the RVDs without a single preferred base according to their total activity level. The abundance of each RVD in natural TALE sequences, as determined using all available *Xanthomonas* TALE sequences in GenBank, is plotted on a log scale (bottom). All bases in the TALE binding site are color-coded (green for A, red for T, orange for G, and blue for C). NLS, nuclear localization signal; VP64, VP64 viral activation domain; 2A, 2A peptide linker; Gluc, *Gaussia* luciferase; minCMV, minimal CMV promoter; Cluc, *Cypridina* luciferase; polyA signal, poly-adenylation signal. All results are collected from three independent experiments in HEK 293FT cells. Error bars indicate s.e.m.; $n = 3$.

The 23 RVD-TALEs exhibited a wide range of DNA base preferences and biological activities in our reporter assay (Fig. 4.1b). In particular, NH- and HN-TALEs activated the guanine-reporter preferentially and at similar levels as the NN-TALE. Interestingly, the NH-TALE also exhibited significantly higher specificity for the G-reporter than the NN-TALE (ratio of G- to A-reporter activations: 16.9 for NH-TALE and 2.7for NN-TALE; Fig. 4.1b), suggesting that NH might be a more optimal RVD for targeting guanines. Our computational analysis of TALE-RVD specificity using a recently published crystal structure of TALE-dsDNA complex(*107, 108*) also suggests that NH has a significantly higher affinity for guanine than NN (Fig. 4.2). We found that substitution of NN with NH in one repeat within the TALE DNA binding domain resulted in a gain of 0.86 ± 0.67 kcal/mol in free energy ($\Delta\Delta G$) in the DNA bound state (Fig. 4.2). This result could be explained by the observation that the imidazole ring on the histidine residue (NH RVD) has a more compact base-stacking interaction with the target guanine base (Fig. 4.2b), indicating that NH would be able to bind guanine more tightly

than NN and thus suggesting a possible mechanism for the increased specificity of NH for guanine. Additionally, the RVD NA exhibited similar levels of reporter activation for all four bases and may be a promising candidate for high efficiency targeting of degenerate DNA sequences in scenarios where non-specific binding is desired(*103*).

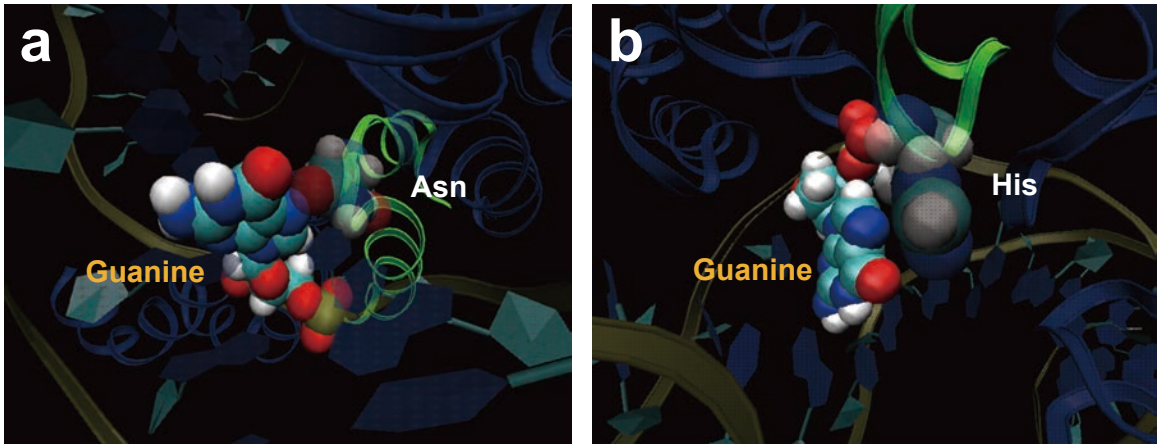


Figure 4.2 Computational analysis of TALE RVD Specificity. We performed extensive free energy perturbation (FEP) calculations for the relative binding affinities between the TALE and its bound DNA. Images show the three-dimensional configuration and results of the free energy calculation for NN:G (**a**) and NH:G (**b**) interactions from one repeat in the TALE-DNA complex. The second amino acid of the guanine-recognizing RVD (i.e., asparagine for RVD NN and histidine for RVD NH) and the guanine base of the bound double-stranded DNA are presented in space filling model and labeled. The free energy calculation results are listed below their corresponding structures.

4.2.2 Relative activity and specificity of guanine-binding RVDs

To determine whether NH and HN are suitable replacements for NN as the G-specific RVD, we directly compared the specificity and activity strength of NN, NK, NH, and HN. We chose two 18bp targets within the *CACNA1C* locus in the human genome and constructed four TALEs for each target, using NN, NK, NH, or HN as the G-targeting RVD (Fig. 4.3a). Since the screening result (Fig. 4.1b) suggested that HN might be less discriminatory than NH when the targeted base is A instead of G, we first designed a luciferase assay to further characterize the G-specificity of each RVD. For each *CACNA1C* target site, we constructed four luciferase reporters: wild type genomic target, and wild type target with 2, 4, or all guanines mutated into adenines (Fig. 4.3a, G-to-A reporters), and compared the activity of each TALE using these reporters (Fig. 4.3a). For both *CACNA1C*target sites, we found that the TALE with NH as the G-targeting RVD exhibited significantly higher specificity for guanine over adenine than the corresponding NK-, HN-, and NN-containing TALEs. For target site 1, introduction of 2 G to A mutations led to 35.4% (TALE1-NN), 40.3% (TALE1-NK), 71.4% (TALE1-NH), and 30.8% (TALE1-HN) of reduction in luciferase activity. For target site 2, two G-to-A mutations led to 21.8% (TALE2-NN), 36.3% (TALE2-NK), 66.1% (TALE2-NH), and 13.9% (TALE2-HN) reduction in reporter activity. Additional G-to-A mutations resulted in further reduction of reporter activity, with NH exhibiting the highest level of discrimination (4a). Additionally, NH TALEs exhibited significantly higher levels of reporter induction than NK TALEs (1.9 times for site 1 and 2.7 times for site 2), and comparable to NN and NH TALEs (Fig. 4.3a). Thus, we decided to focus on the RVDs

NN, NK, and NH in subsequent experiments to assess their usefulness in modulating transcription at endogenous human genome targets.

Figure 4.3 Characterization of guanine-specific repeat-variable diresidues (RVDs). a, specificity and activity of different Guanine-targeting RVDs. Schematic showing the selection of two TALE binding sites within the *CACNA1C* locus of the human genome. The TALE RVDs are shown above the binding site sequences and yellow rectangles indicate positions of G-targeting RVDs (left). Four different TALEs using NN, NK, NH, and HN as the putative G-targeting RVD were synthesized for each target site. The specificity for each putative G-targeting RVD is assessed using luciferase reporter assay, by measuring the levels of reporter activation of the wild-type TALE binding site and mutant binding sites, with either 2, 4, or all guanines substituted by adenine. The mutated guanines and adenines are highlighted with orange and green respectively. **b,** Endogenous transcriptional modulation using TALEs containing putative G-specific RVDs. TALEs using NN, NK, NH, and HN as the G-targeting RVD were synthesized to

(Figure 4.3, continued) target two distinct 18bp target sites in the human *CACNA1C* locus. Changes in mRNA are measured using qRT-PCR as described previously(*27*). VP64, VP64 transcription activation domain. All results are collected from three independent experiments in HEK 293FT cells. Error bars indicate s.e.m.; $n = 3$.

4.2.3 Evaluation of guanine-binding RVDs at endogenous genome loci

Using qRT-PCR, we further compared the performance of NN, NK, NH, and HN for targeting endogenous genomic sequences. We tested the ability of NN-, NK-, NH-, and HN- TALEs to activate *CACNA1C* transcription by targeting the two endogenous target sites (Fig. 4.3b). To control for differences in TALE expression levels, all TALE were fused to 2A-GFP and exhibited similar levels of GFP fluorescence(*27*). Using qRT-PCR, we found that the endogenous activity of each TALE corresponded to the reporter assay. Both TALE1-NH and TALE2-NH were able to achieve similar levels of transcriptional activation as TALE1-NN and TALE2-NN (~5 and ~3 folds of activation for targets 1 and 2 respectively) and twice more than TALE1-NK and TALE2-NK (Fig. 4.3b). Although TALE1-HN and TALE2-HN exhibited comparable activity with TALEs bearing RVDs NN and NH, the lack of specificity in distinguishing guanine and adenosine bases as shown in previous test (Fig. 4.3a) does not warrant the superiority of HN over existing guanine-binding RVDs. On the other hand, based on all the results from specificity and endogenous activity tests, the RVD NH seems to be a more suitable substitute for NN than NK when higher targeting specificity is desired, as it also provides higher levels of biological activity. Further testing using additional endogenous genomic targets will help validate the broad utility of NH as a highly specific G-targeting RVD.

4.2.4 Development of mammalian TALE transcriptional repressors

Having identified NH as a more specific G-recognizing RVD, we sought to develop a mammalian TALE repressor architecture to enable researchers to suppress transcription of endogenous genes. TALE repressors have the potential to suppress the expression of genes as well as non-coding transcripts such as microRNAs, rendering it a highly desirable tool for testing the causal role of specific genetic elements. In order to identify a suitable repression domain for use with TALEs in mammalian cells, we used a TALE targeting the promoter of the human *SOX2* gene to evaluate the transcriptional repression activity of a collection of candidate repression domains (Fig. 4.4a). We selected repression domains across a range of eukaryotic host species to increase the change of finding a potent synthetic repressor, including the PIE-1repression domain (PIE-1)(*109*) from *Caenorhabditis elegans*, the QA domain within the Ubx gene (Ubx-QA)(*110*) from *Drosophila melanogaster*, the IAA28 repression domain (IAA28-RD)(*111*) from *Arabidopsis thaliana*, the mSin interaction domain (SID)(*106*), Tbx3 repression domain (Tbx3-RD), and the Krüppel associated box (KRAB)(*112*) repression domain from *Homo Sapiens* (Fig. 4.4b). Since different truncations of KRAB have been known to exhibit varying levels of transcriptional repression(*112*), we tested three different truncations of KRAB (Fig. 4.4c). We expressed these candidate TALE repressors in HEK 293FTcells and found that TALEs carrying two widely used mammalian transcriptional repression domains, the SID(*106*) and KRAB(*112*) domains, were able to repress endogenous *SOX2* expression, while the other domains had little effect on transcriptional activity (Fig. 4.4c). To control for potential perturbation of SOX2 transcription due to TALE binding, expression of the SOX2-targeting TALE DNA binding domain alone without any

effector domain had no effect (similar to mock or expression of GFP) on the transcriptional activity of SOX2 (Fig. 4.4c, Null condition). Since the SID domain was able to achieve 26% more transcriptional repression of the endogenous *SOX2* locus than the KRAB domain (Fig. 4.4c), we decided to use the SID domain for our subsequent studies.

To further test the effectiveness of the SID repressor domain for down regulating endogenous transcription, we combined SID with CACNA1C-target TALEs from the previous experiment (Fig. 4.3, Fig. 4.4d). Using qRT-PCR, we found that replacement of the VP64 domain on CACNA1C-targeting TALEs with SID was able to repress CACNA1C transcription. Additionally, similar to the transcriptional activation study (Fig. 4.3b, left), NH-containing TALE repressor was able to achieve a similar level of transcriptional repression as the NN-containing TALE (~4 fold repression), while the TALE repressor using NK was significantly less active (~2 fold repression) (Fig. 4.4d). These data demonstrate that SID is indeed a suitable repression domain, while also further supporting NH as a more suitable G-targeting RVD than NK.

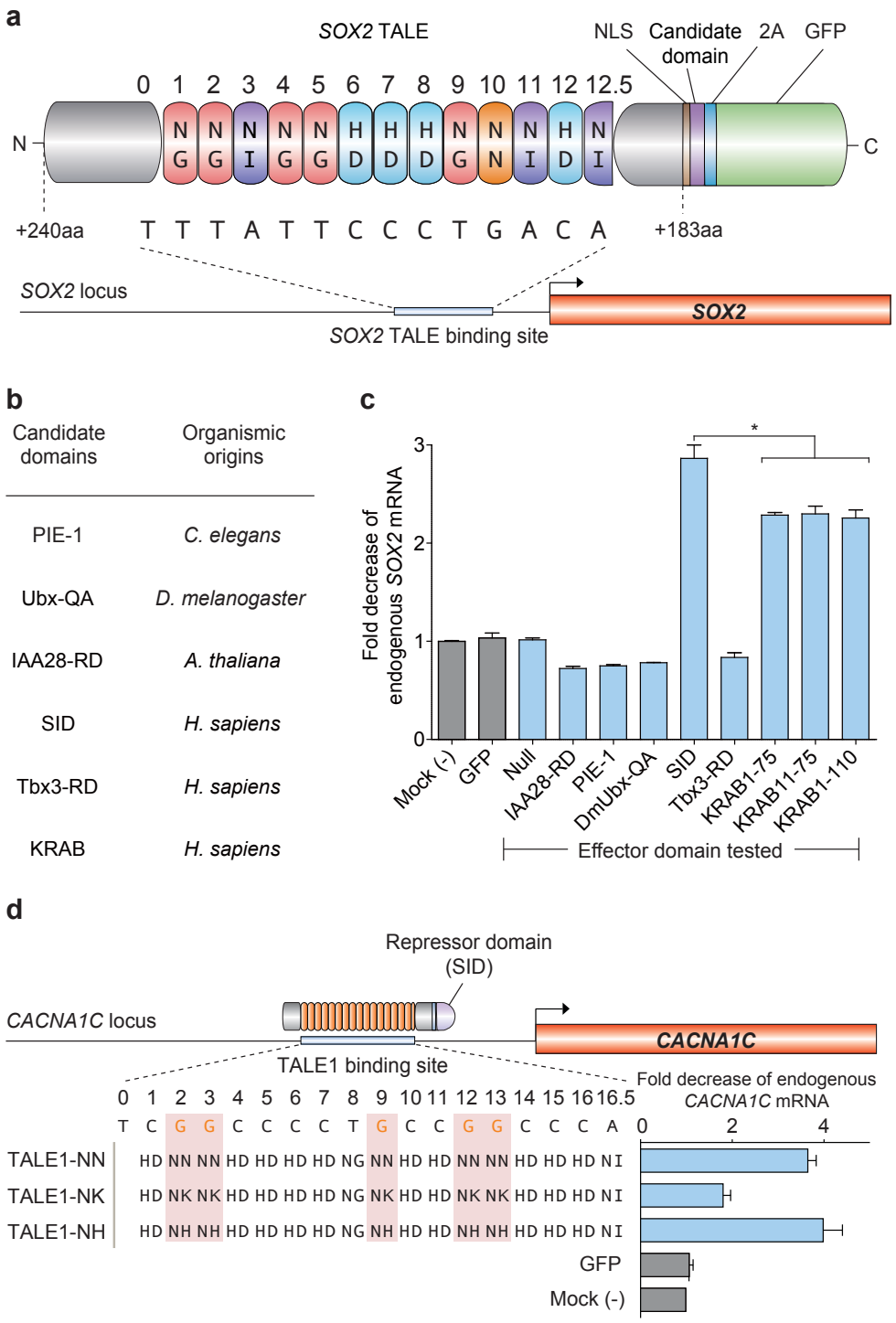


Figure 4.4 Development of aTALE transcriptional repressor architecture. a, Design of *SOX2* TALE for TALE repressor screening. A TALE targeting a 14bp sequence within the *SOX2* locus of the human genome was synthesized as described previously(*27*). **b,** List of all repressors screened and their host origin (left). Eight different candidate

(Figure 4.4, continued) repressor domains were fused to the C-term of the *SOX2* TALE. **c,** The fold decrease of endogenous SOX2 mRNA is measured using qRT-PCR by dividing the *SOX2* mRNA levels in mock transfected cells by *SOX2* mRNA levels in cells transfected with each candidate TALE repressor. **d,** Transcriptional repression of endogenous CACNA1C. TALEs using NN, NK, and NH as the G-targeting RVD were constructed to target a 18bp target site within the human *CACNA1C* locus (site 1 in Figure 4.2). Each TALE is fused to the SID repression domain. NLS, nuclear localization signal; KRAB, Krüppel-associated box; SID, mSin interaction domain. All results are collected from three independent experiments in HEK 293FT cells. Error bars indicate s.e.m.; $n = 3$. * $p < 0.05$, Student's t test.

4.3 Discussion

TALEs can be easily customized to recognize specific sequences on the endogenous genome. Here, we conducted a series of screens to address two important limitations of the TALE toolbox. Together, the identification of a more stringent G-specific RVD with uncompromised activity strength as well as a robust TALE repressor architecture further expands the utility of TALEs for probing mammalian transcription and genome function.

4.4 Methods

4.4.1 Construction of TALE activators, repressors and reporters

All TALE activators or repressors were constructed as previously described using a hierarchical ligation strategy(*27*). The sequences for all constructs used in this study can be found in Supplementary Table 1. To control for differences in the expression of each

TALE, all TALEs are in-frame fused with the *Gaussia* luciferase (Gluc) gene via a 2A linker. The Gluc gene will be translated in an equimolar amount as TALEs. Truncation variants of the Krüppel-associated box (KRAB) domain, the PIE-1 repression domain (PIE-1), the QA domain within the Ubx gene (Ubx-QA), the IAA28 repression domain (IAA28-RD), Tbx3 repression domain (Tbx3-RD), and the mSin interaction domain (SID) were codon optimized for mammalian expression and synthesized with flanking *Nhe*I and *Xba*I restriction sites (Genscript). All repressor domains were cloned into the TALE backbone by replacing the VP64 activation domain using *Nhe*I and *Xba*I restriction sites. To control for any effect on transcription resulting from TALE binding, we constructed expression vectors carrying the TALE DNA binding domain alone using PCR cloning. The coding regions of all constructs were completely verified using Sanger sequencing.

All luciferase reporter plasmids were designed and synthesized by inserting the TALE binding site upstream of the minimal CMV promoter driving the expression of a *Cypridina* luciferase (Cluc) gene (Fig. 4.1), similar to minCMV-mCherry reporter used in previous studies(*27*).

4.4.2 Cell culture and luciferase reporter activation assay

Maintenance of human embryonic kidney cell line HEK 293FT (Invitrogen) were carried out with Dulbecco's modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum (HyClone), 2mM GlutaMAX (Invitrogen), 100U/mL Penicillin, and 100μg/mL Streptomycin, under 37ºC, 5% CO_2 incubation condition.

Luciferase reporter assays were performed by co-transfecting HEK 293FT cells with TALE-2A-luciferaseexpression and luciferase reporter plasmids. In the case of the reporter-only control, cells were co-transfected with a control *Gaussia* luciferase plasmid (pCMV-Gluc, New England BioLabs). HEK 293FT cells were seeded into 24-well plates the day prior to transfection at densities of $2x10^5$ cells/well. Approximately 24h after initial seeding, cells were transfected using Lipofectamine2000 (Invitrogen) following the manufacturer's protocol. For each well of the 24-well plates 700ng of dTALE and 50ng of each reporter plasmids were used to transfect HEK 293FT cells.

Dual luciferase reporter assays were carried out with the BioLux *Gaussia* luciferase flex assay kit and BioLux *Cypridina* luciferase assay kit (New England Biolabs) following the manufacturer's recommended protocol. Briefly, media from each well of transfected cells were collected 48 hours after transfection. For each sample, 20uL of the media were added into a 96-well assay plate, mixed with each one of the dual luciferase assay mixes. After briefly incubation, as indicated in the manufacturer's protocol, luminescence levels of each sample were measured using the Varioskan flash multimode reader (Thermo Scientific).

The activity of each TALE is determined by measuring the level of luciferase reporter induction, calculated as the level of Cluc induction in the presence of TALE activator minus the level of Cluc induction without TALE activator. The activity of each TALE is normalized to the level of TALE expression as determined by the Gluc activity level

(each TALE is in frame fused to 2A-Gluc), to control for differences in cell number, sample preparation, transfection efficiency, and protein expression level. The concentration of all DNA used in transfection experiments were determined using gel analysis.

We determined the base preference of each RVD according to the induction of each base-specific reporters by the corresponding RVD screening TALE (RVD-TALE, Figure 4.1a). Statistical analysis were performed using one-way analysis of variance (ANOVA) tests. Each RVD was tested by taking the reporter with the highest luciferase activity as the putative preferred base and comparing it with the remaining three bases as a group. For a given RVD, if the putative preferred base gave statistically significant test results ($p < 0.05$, one-way ANOVA), we classified that RVD as having a single preferred base, otherwise that RVD is tagged as not having a single preferred base.

4.4.3 Endogenous gene transcriptional activation assay

For the endogenous gene transcriptional level assay to test the biological activities of TALE activators and TALE repressors, HEK 293FT cells were seeded into 24-well plates. 1ug of TALE plasmid was transfected using Lipofectamine2000 (Invitrogen) according to manufacturer's protocol. Transfected cells were cultured at 37°C for 72 hours before RNA extraction. At least 100,000 cells were harvested and subsequently processed for total RNA extraction using the RNAeasyPlus Mini Kit (Qiagen). cDNA was generated using the High Capacity RNA-to-cDNA Master Mix (Applied Biosystems) according to the manufacturer's recommended protocol. After cDNA synthesis, cDNA

from each samples were added to the qRT-PCR assay with the Taqman Advanced PCR Master Mix (Applied Biosystems) using a StepOne Plus qRT-PCR machine. The fold activation in the transcriptional levels of SOX2 and CACNA1C mRNA were detected using standard TaqMan Gene Expression Assays with probes having the best coverage (Applied Biosystems; *SOX2*, Hs01053049_s1; *CACNA1C*, Hs00167681_m1).

4.4.4 Computational analysis of RVD specificity

To assess the guanine-specificity of NH, we performed extensive computational simulations to compare the relative binding affinities between guanine and NN or NH using free energy perturbation (FEP)(*113, 114*), a widely used approach for calculating binding affinities for a variety of biological interactions, such as ligand-receptor binding, protein–protein interaction, and protein-nucleic acid binding(*115, 116*). Molecular dynamics simulations were carried out as previously described(*115, 116*). We based our calculations on the recently released crystal structure of the TALE PthXo1 bound to DNA (PDB ID: 3UGM)(*107*). We used a fragment of the crystal structure containing repeats 11-18 of PthXo1 (RVD sequence: HD[11]-NG[12]-NI[13]-HD[14]-NG[15]-NN[16]-NG[17]-NI[18], repeat number specified in square brackets) and the corresponding double-stranded DNA molecule containing the TALE binding sequence (5'-CTACTGTA-3') to compare the binding affinities of RVDs NN, NK, and NH for guanine. Since the 16th repeat in the structure is NN, we computationally mutated NN into NH or NK and calculated the binding affinity of each configuration (NN:G, NH:G). The affinity was calculated as the gain of free energy ($\Delta\Delta G$) in the DNA bound state taking NN:G as reference ($\Delta\Delta G = 0$).

Page intentionally left blank.

5. Multiplex Genome Engineering Using CRISPR/Cas Systems

The work described in this chapter is done with Fei Ann Ran as equal contributors, and has been published as Cong L*, Ran FA*, et al. Science. 2013 (*33*).

5.1 Introduction

Precise and efficient genome targeting technologies are needed to enable systematic reverse engineering of causal genetic variations by allowing selective perturbation of individual genetic elements. Although genome-editing technologies such as designer zinc fingers (ZFs) (*15, 27, 60, 117*), transcription activator-like effectors (TALEs) (*23, 25, 27, 39, 53, 60, 118*), and homing meganucleases (*119*) have begun to enable targeted genome modifications, there remains a need for new technologies that are scalable, affordable, and easy to engineer. Here, we report the development of a new class of precision genome engineering tools based on the RNA-guided Cas9 nuclease (*30, 120, 121*) from the type II prokaryotic CRISPR adaptive immune system (*31, 32, 122, 123*).

The *Streptococcus pyogenes* SF370 type II CRISPR locus consists of four genes, including the Cas9 nuclease, as well as two non-coding RNAs: tracrRNA and a pre-crRNA array containing nuclease guide sequences (spacers) interspaced by identical direct repeats (DRs) (Fig. S1) (*124*). We sought to harness this prokaryotic RNA-programmable nuclease system to introduce targeted double stranded breaks (DSBs) in mammalian chromosomes through heterologous expression of the key components. It has been previously shown that expression of tracrRNA, pre-crRNA, host factor RNase III,

and Cas9 nuclease are necessary and sufficient for cleavage of DNA *in vitro* (*120, 121*) and in prokaryotic cells (*125, 126*). We codon optimized the *S. pyogenes Cas9 (SpCas9)* and *RNase III (SpRNase III)* and attached nuclear localization signals (NLS) to ensure nuclear compartmentalization in mammalian cells. Expression of these constructs in human 293FT cells revealed that two NLSs are most efficient at targeting SpCas9 to the nucleus (Fig. 5.1A). To reconstitute the non-coding RNA components of CRISPR, we expressed an 89-nucleotide (nt) tracrRNA (Fig. S2) under the RNA polymerase III U6 promoter (Fig. 5.1B). Similarly, we used the U6 promoter to drive the expression of a pre-crRNA array comprising a single guide spacer flanked by DRs (Fig. 5.1B). We designed our initial spacer to target a 30-basepair (bp) site (protospacer) in the human *EMX1* locus that precedes an NGG, the requisite protospacer adjacent motif (PAM) (Fig. 5.1C and fig. S1) (*127, 128*).

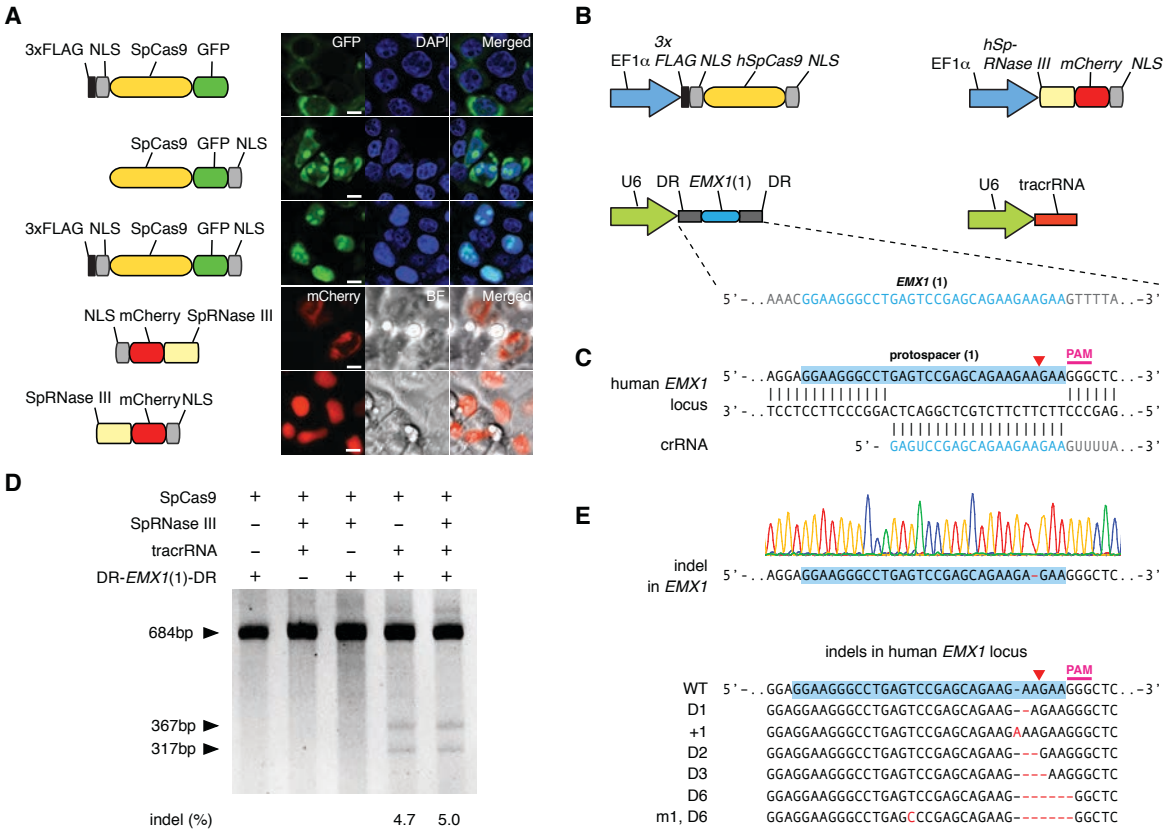


Figure 5.1 The Type II CRISPR locus from *Streptococcus pyogenes* SF370 can be reconstituted in mammalian cells to facilitate targeted DSBs of DNA. (**A**) Engineering of SpCas9 and SpRNase III with NLSs enables import into the mammalian nucleus. (**B**) Mammalian expression of SpCas9 and SpRNase III are driven by the EF1a promoter, whereas tracrRNA and pre-crRNA array (DR-Spacer-DR) are driven by the U6 promoter. A protospacer (blue highlight) from the human *EMX1* locus with PAM is used as template for the spacer in the pre-crRNA array. (**C**) Schematic representation of base pairing between target locus and *EMX1*-targeting crRNA. Red arrow indicates putative cleavage site. (**D**) SURVEYOR assay for SpCas9-mediated indels. (**E**) An example chromatogram showing a micro-deletion, as well as representative sequences of mutated alleles identified from 187 clonal amplicons. Red dashes, deleted bases; red bases, insertions or mutations. Scale bar = 10μm.

5.2 Reconstitution of the CRISPR/Cas system in mammalian cells

To test whether heterologous expression of the CRISPR system (SpCas9, SpRNase III, tracrRNA, and pre-crRNA) can achieve targeted cleavage of mammalian chromosomes, we transfected 293FT cells with different combinations of CRISPR components. Since DSBs in mammalian DNA are partially repaired by the indel-forming non-homologous end joining (NHEJ) pathway, we used the SURVEYOR assay (Fig. S3) to detect endogenous target cleavage (Fig. 5.1D and fig. S2B). Co-transfection of all four required CRISPR components resulted in efficient cleavage of the protospacer (Fig. 5.1D and fig. S2B), which is subsequently verified by Sanger sequencing (Fig. 5.1E). Interestingly, SpRNase III was not necessary for cleavage of the protospacer (Fig. 5.1D), and the 89-nt tracrRNA is processed in its absence (Fig. S2C). Similarly, maturation of pre-crRNA does not require RNase III (Fig. 5.1D and fig. S4), suggesting that there may be endogenous mammalian RNases that assist in pre-crRNA maturation (*129-131*). Removing any of the remaining RNA or Cas9 components abolished the genome cleavage activity of the CRISPR system (Fig. 5.1D). These results define a minimal three-component system for efficient CRISPR-mediated genome modification in mammalian cells.

5.3 Endogenous genome cleavage by CRISPR/Cas system

Next, we explored the generalizability of CRISPR-mediated cleavage in eukaryotic cells by targeting additional protospacers within the *EMX1* locus (Fig. 5.2A). To improve co-delivery, we designed an expression vector to drive both pre-crRNA and SpCas9 (Fig. S5). In parallel, we adapted a chimeric crRNA-tracrRNA hybrid (Fig. 5.2B, top) design

recently validated *in vitro* (*120*), where a mature crRNA is fused to a partial tracrRNA via a synthetic stem-loop to mimic the natural crRNA:tracrRNA duplex (Fig. 5.2B, bottom). We observed cleavage of all protospacer targets when SpCas9 is co-expressed with pre-crRNA (DR-spacer-DR) and tracrRNA. However, not all chimeric RNA designs could facilitate cleavage of their genomic targets (Fig. 5.2C, Table S1). We then tested targeting of additional genomic loci in both human and mouse cells by designing pre-crRNAs and chimeric RNAs targeting the human *PVALB* and the mouse *Th* loci (Fig. S6). We achieved efficient modification at all three mouse *Th* and one *PVALB* targets using the crRNA:tracrRNA design, thus demonstrating the broad applicability of the CRISPR system in modifying different loci across multiple organisms (Table S1). For the same protospacer targets, cleavage efficiencies of chimeric RNAs were either lower than those of crRNA:tracrRNA duplexes or undetectable. This may be due to differences in the expression and stability of RNAs, degradation by endogenous RNAi machinery, or secondary structures leading to inefficient Cas9 loading or target recognition.

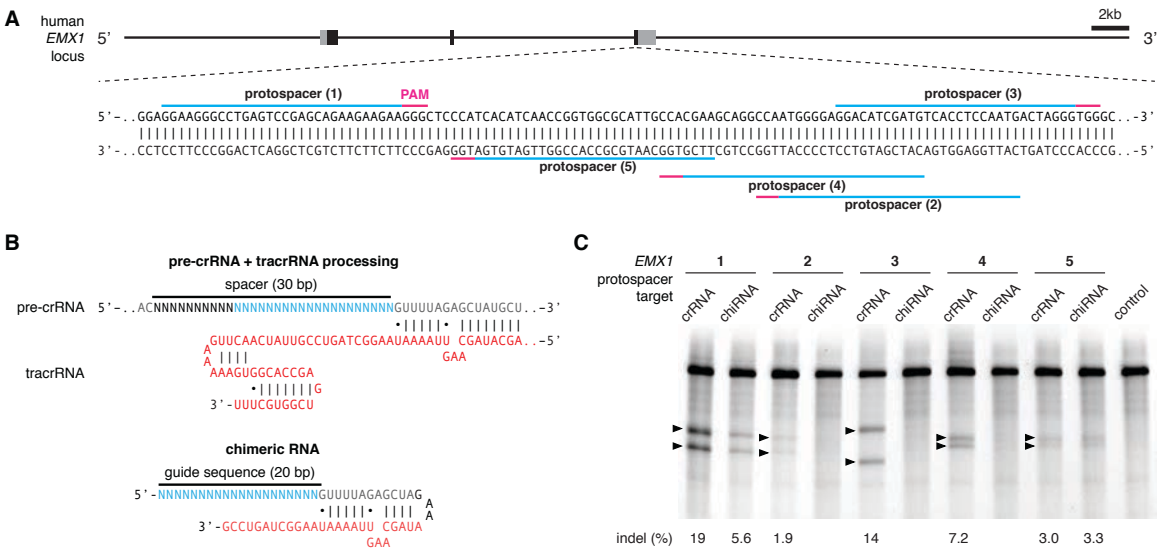


Figure 5.2 SpCas9 can be reprogrammed to target multiple genomic loci in mammalian cells. (**A**) Schematic of the human *EMX1* locus showing the location of five protospacers, indicated by blue lines with corresponding PAM in magenta. (**B**) Schematic of the pre-crRNA:tracrRNA complex (top) showing hybridization between the direct repeat (gray) region of the pre-crRNA and tracrRNA. Schematic of a chimeric RNA design (*120*) (bottom). tracrRNA sequence is shown in red and the 20bp spacer sequence in blue. (**C**) SURVEYOR assay comparing the efficacy of Cas9-mediated cleavage at five protospacers in the human *EMX1* locus. Each protospacer is targeted using either processed pre-crRNA:tracrRNA complex (crRNA) or chimeric RNA (chiRNA).

5.4 Target cleavage specificity of the CRISPR/Cas system in mammalian cells

Effective genome editing requires that nucleases target specific genomic loci with both high precision and efficiency. To investigate the specificity of CRISPR-mediated cleavage, we analyzed single-nucleotide mismatches between the spacer and its mammalian protospacer target (Fig. 5.3A). We observed that single-base mismatch up to 12-bp 5' of the PAM completely abolished genomic cleavage by SpCas9, whereas

spacers with mutations farther upstream retained activity against the protospacer target (Fig. 5.3B). This is consistent with previous bacterial and *in vitro* studies of Cas9 specificity (*120, 125*). Furthermore, CRISPR is able to mediate genomic cleavage as efficiently as a pair of TALE nucleases (TALEN) targeting the same *EMX1* protospacer (Fig. 5.3, C and D).

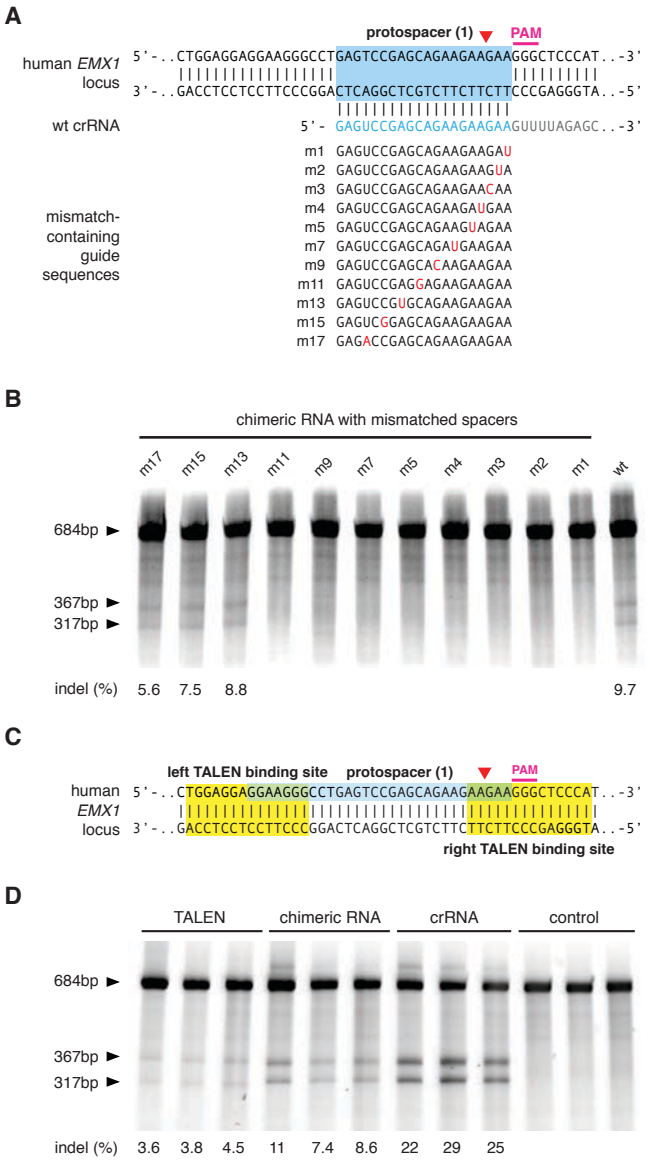


Figure 5.3 Evaluation of the SpCas9 specificity and comparison of efficiency with TALENs. (**A**) *EMX1*-targeting chimeric crRNAs with single point mutations were generated to evaluate the effects of spacer-protospacer mismatches. (**B**) SURVEYOR assay comparing the cleavage efficiency of different mutant chimeric RNAs. (**C**) Schematic showing the design of TALENs targeting *EMX1*. (**D**) SURVEYOR gel comparing the efficiency of TALEN and SpCas9 ($N = 3$).

5.5 Development of a Cas9 nickase and its application in inducing homology-directed repair

Targeted modification of genomes ideally avoids mutations arising from the error-prone NHEJ mechanism. The wild-type SpCas9 is able to mediate site-specific DSBs, which can be repaired through either NHEJ or homology-directed repair (HDR). We engineered an aspartate-to-alanine substitution (D10A) in the RuvC I domain of SpCas9 to convert the nuclease into a DNA nickase (SpCas9n, Fig. 5.4A) (*120, 121, 125*), because nicked genomic DNA is typically repaired either seamlessly or through high-fidelity HDR. SURVEYOR (Fig. 5.4B) and sequencing of 327 amplicons did not detect any indels induced by SpCas9n. However, it is worth noting that nicked DNA can in rare cases be processed via a DSB intermediate and result in a NHEJ event (*132*). We then tested Cas9-mediated HDR at the same *EMX1* locus with a homology repair template to introduce a pair of restriction sites near the protospacer (Fig. 5.4C). SpCas9 and SpCas9n catalyzed integration of the repair template into *EMX1* locus at similar levels (Fig. 5.4D), which we further verified via Sanger sequencing (Fig. 5.4E). These results demonstrate the utility of CRISPR for facilitating targeted genomic insertions. Given the 14-bp (12-bp from the seed sequence and 2-bp from PAM) target specificity (Fig. 5.3B) of the wild type SpCas9, the use of a nickase may reduce off-target mutations.

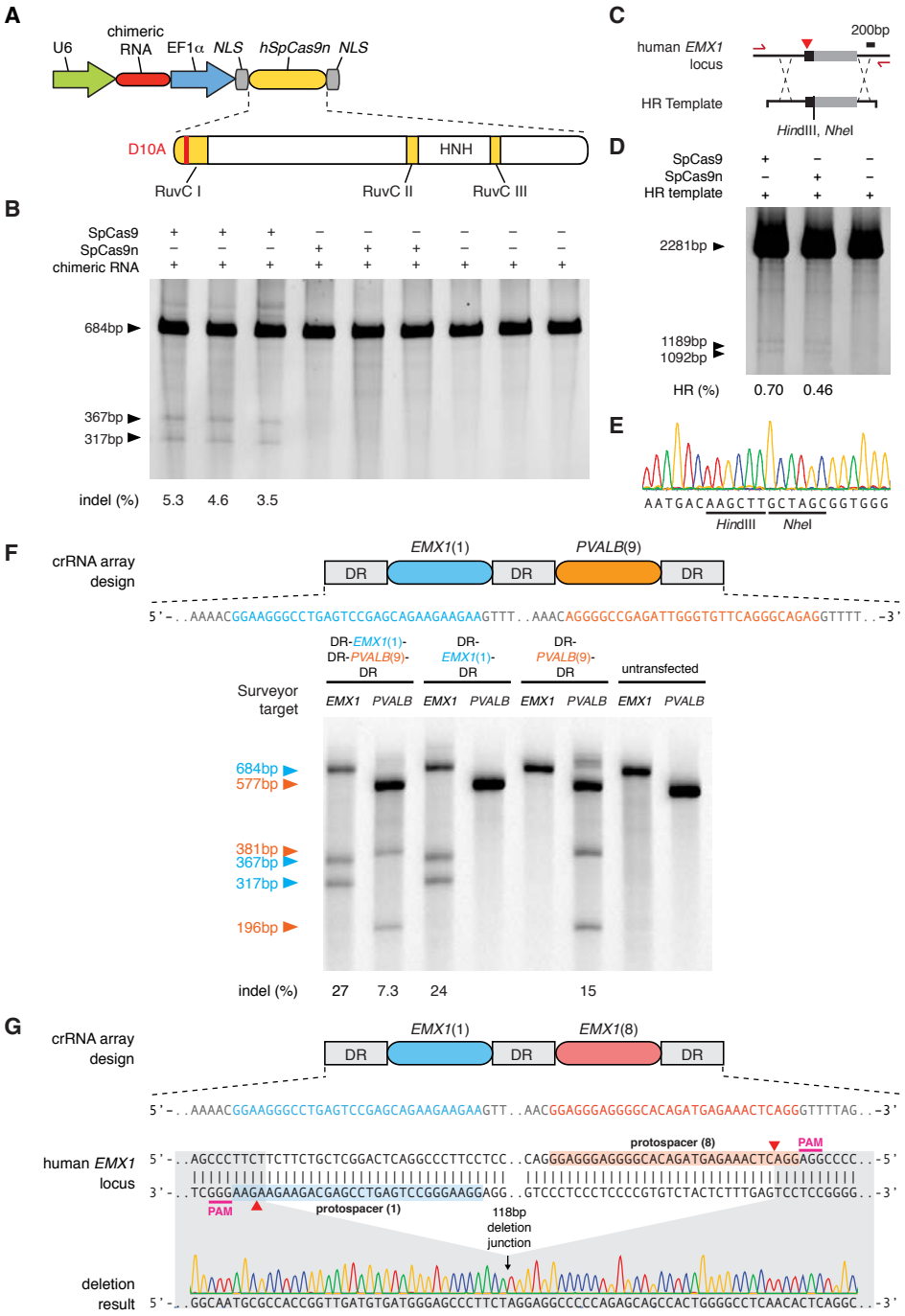


Figure 5.4 Applications of Cas9 for homologous recombination and multiplex genome engineering. (**A**) Mutation of the RuvC I domain converts Cas9 into a nicking enzyme (SpCas9n) (**B**) Co-expression of *EMX1*-targeting chimeric RNA with SpCas9 leads to indels, whereas SpCas9n does not ($N = 3$). (**C**) Schematic representation of the recombination strategy. A repair template is designed to insert restriction sites into *EMX1* locus. Primers used to amplify the modified region are shown as red arrows. (**D**)

(Figure 5.4, continued) Restriction fragments length polymorphism gel analysis. Arrows indicate fragments generated by *Hin*dIII digestion. (**E**) Example chromatogram showing successful recombination. (**F**) SpCas9 can facilitate multiplex genome modification using a crRNA array containing two spacers targeting *EMX1* and *PVALB*. Schematic showing the design of the crRNA array (top). Both spacers mediate efficient protospacer cleavage (bottom). (**G**) SpCas9 can be used to achieve precise genomic deletion. Two spacers targeting *EMX1* (top) mediated a 118bp genomic deletion (bottom).

5.6 Multiplexed mammalian genome engineering with CRISPR/Cas system

Finally, the natural architecture of CRISPR loci with arrayed spacers (Fig. S1) suggests the possibility of multiplexed genome engineering. Using a single CRISPR array encoding a pair of *EMX1*- and *PVALB*-targeting spacers, we detected efficient cleavage at both loci (Fig. 5.4F). We further tested targeted deletion of larger genomic regions through concurrent DSBs using spacers against two targets within *EMX1* spaced by 119-bp, and observed a 1.6% deletion efficacy (3 out of 182 amplicons; Fig. 5.4G), thus demonstrating the CRISPR system can mediate multiplexed editing within a single genome.

5.7 Potential of CRISPR/Cas systems for genome engineering

The ability to use RNA to program sequence-specific DNA cleavage defines a new class of genome engineering tools. Here, we have shown that the *S. pyogenes* CRISPR system can be heterologously reconstituted in mammalian cells to facilitate efficient genome editing; an accompanying study has independently confirmed high efficiency CRISPR-

mediated genome targeting in several human cell lines (*133*). However, several aspects of the CRISPR system can be further improved to increase its efficiency and versatility. The requirement for an NGG PAM restricts the *S. pyogenes* CRISPR target space to every 8-bp on average in the human genome (Fig. S7), not accounting for potential constraints posed by crRNA secondary structure or genomic accessibility due to chromatin and DNA methylation states. Some of these restrictions may be overcome by exploiting the family of Cas9 enzymes and its differing PAM requirements (*127, 128*) across the microbial diversity (*122*). Indeed, other CRISPR loci are likely to be transplantable into mammalian cells; for example, the *Streptococcus thermophilus* LMD-9 CRISPR1 can also mediate mammalian genome cleavage (Fig. S8). Finally, the ability to carry out multiplex genome editing in mammalian cells enables powerful applications across basic science, biotechnology, and medicine (*134*).

5.8 Materials and methods

5.8.1 Cell culture and transfection

Human embryonic kidney (HEK) cell line 293FT (Life Technologies) was maintained in Dulbecco's modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum (HyClone), 2mM GlutaMAX (Life Technologies), 100U/mL penicillin, and 100μg/mL streptomycin at 37ºC with 5% CO_2 incubation. Mouse neuro2A (N2A) cell line (ATCC) was maintained with DMEM supplemented with 5% fetal bovine serum (HyClone), 2mM GlutaMAX (Life Technologies), 100U/mL penicillin, and 100μg/mL streptomycin at 37ºC with 5% CO_2.

293FT or N2A cells were seeded into 24-well plates (Corning) one day prior to transfection at a density of 200,000 cells per well. Cells were transfected using Lipofectamine 2000 (Life Technologies) following the manufacturer's recommended protocol. For each well of a 24-well plate a total of 800ng plasmids was used.

5.8.2 Surveryor assay and sequencing analysis for genome modification

293FT or N2A cells were transfected with plasmid DNA as described above. Cells were incubated at 37ºC for 72 hours post transfection before genomic DNA extraction. Genomic DNA was extracted using the QuickExtract DNA extraction kit (Epicentre) following the manufacturer's protocol. Briefly, cells were resuspended in QuickExtract solution and incubated at 65ºC for 15 minutes and 98ºC for 10 minutes.

Genomic region surrounding the CRISPR target site for each gene was PCR amplified, and products were purified using QiaQuick Spin Column (Qiagen) following manufacturer's protocol. A total of 400ng of the purified PCR products were mixed with 2μl 10X Taq polymerase PCR buffer (Enzymatics) and ultrapure water to a final volume of 20μl, and subjected to a re-annealing process to enable heteroduplex formation: 95ºC for 10min, 95ºC to 85ºC ramping at – 2ºC/s, 85ºC to 25ºC at – 0.25ºC/s, and 25ºC hold for 1 minute. After re-annealing, products were treated with SURVEYOR nuclease and SURVEYOR enhancer S (Transgenomics) following the manufacturer's recommended protocol, and analyzed on 4-20% Novex TBE poly-acrylamide gels (Life Technologies). Gels were stained with SYBR Gold DNA stain (Life Technologies) for 30 minutes and

imaged with a Gel Doc gel imaging system (Bio-rad). Quantification was based on relative band intensities.

5.8.3 Restriction fragment length polymorphism assay for detection of homologous recombination

HEK 293FT and N2A cells were transfected with plasmid DNA, and incubated at 37°C for 72 hours before genomic DNA extraction as described above. The target genomic region was PCR amplified using primers outside the homology arms of the homologous recombination (HR) template. PCR products were separated on a 1% agarose gel and extracted with MinElute GelExtraction Kit (Qiagen). Purified products were digested with *Hin*dIII (Fermentas) and analyzed on a 6% Novex TBE poly-acrylamide gel (Life Technologies).

5.8.4 RNA extraction and purification

HEK 293FT cells were maintained and transfected as stated previously. Cells were harvested by trypsinization followed by washing in phosphate buffered saline (PBS). Total cell RNA was extracted with TRI reagent (Sigma) following manufacturer's protocol. Extracted total RNA was quantified using Naonodrop (Thermo Scientific) and normalized to same concentration.

5.8.5 Northern blot analysis of small RNA expression in mammalian cells

RNAs were mixed with equal volumes of 2X loading buffer (Ambion), heated to 95ºC for 5 min, chilled on ice for 1 min and then loaded onto 8% denaturing polyacrylamide gels (SequaGel, National Diagnostics) after pre-running the gel for at least 30 minutes. The samples were electrophoresed for 1.5 hours at 40W limit. Afterwards, the RNA was transferred to Hybond N+ membrane (GE Healthcare) at 300 mA in a semi-dry transfer apparatus (Bio-rad) at room temperature for 1.5 hours. The RNA was crosslinked to the membrane using autocrosslink button on Stratagene UV Crosslinker the Stratalinker (Stratagene). The membrane was pre-hybridized in ULTRAhyb-Oligo Hybridization Buffer (Ambion) for 30 min with rotation at 42ºC and then probes were added and hybridized overnight. Probes were ordered from IDT and labeled with [gamma-32P] ATP (Perkin Elmer) with T4 polynucleotide kinase (New England Biolabs). The membrane was washed once with pre-warmed (42ºC) 2xSSC, 0.5% SDS for 1 min followed by two 30 minute washes at 42ºC. The membrane was exposed to phosphor screen for one hour or overnight at room temperature and then scanned with phosphorimager (Typhoon).

6. Conclusion And Future Directions

6.1 Broad implication of the development of genome engineering technologies

The development of novel technologies have always been a driving force for biological and biomedical research, to name just a few of them, the advances of microscopy that transformed classic biology through breaking our observational barriers, the invention of polymerase chain reaction (PCR) that laid the defining pedestals for the modern molecular biology, and the realization of massive parallel sequencing that revolutionized the way we collect, analyze, and apply biological data. In a word, the pursuit of better tools for answering basic questions shapes the way we are discovering and practicing essential principles.

Resounding to this theme, in recent years, the fascinating evolvement of biotechnology, especially in the genome engineering field that enables exact, efficient control of biological systems, has been moving forward beyond our anticipations. It shows no signs of slowing down, continuing to demonstrate its influence at an even accelerated speed. This emerging trend is both driven by the fundamental need of a versatile technological platform to address the age of 'big data' in biology and medicine, and, at the same time, made possible in large part by the bioengineering capacity built upon the realization of genome sequencing and genome-wide data acquisition and analysis. Hence, a key conclusion from this process is that this relationship between the two major technological wave fronts are reciprocal.

The path of my thesis research has been revolving around this point and in fact serves as repeated demonstration of this interplay between observational and interventional technologies. The paradigm-changing surge in our capacity to sequence metagenomic space, largely from the implementation of human genome project and associated initiatives, in joint force with the improved ability to study the function of genes and genome structures, led to the discoveries of TALEs and CRISPR/Cas systems from plant bacteria and various strains of bacteria/archaea, respectively; the engineering of designer TALEs and customizable CRISPR/Cas technologies in return has the potential of transforming our way to address the challenge of making sense of the data generated by genomic sequencing and other investigations, again not alone, but through combining them with mighty read-out methodologies.

This inspires another encouraging development, the rising impact of synthetic biology. Specifically, the integration of all the biology and technologies mentioned above in innovative ways would bring together a cornucopia of synthetic tools that utilizes our biological understandings to re-invent the discipline itself and create incredibly enduring effects upon human health. Therefore I would like to propose a few integrative ideas that might lead to such synthetic biotechnologies and their applications.

6.2 The application of inducible domains to modulate protein activity for temporally and spatially precise control of genome engineering tools

The fulfillment of high-efficiency genome engineering have long been confounded by the issue of accuracy, i.e., off-target effects. To develop reliable disease models or therapeutics with these technologies, the ability to have fine control of these tools is necessary and critical for its applications in the study of human disorders.

Among the various control methodology available for protein activity, the best mechanism that currently allow for fine temporal resolution, low toxicity, and high sensitivity, is light or small molecule-mediated chemical activation and inactivation of protein function (*135-144*). Both directions will be a key component in the promise of developing models and therapeutic interventions for human diseases through application of genome engineering technology.

It is desirable, for example, when creating a complex hierarchical structure of a synthetic circuit, one could control the activity of each circuit parts via optical or chemical stimuli serving as inputs or triggers. In another case, during the production of bio-material through metabolic engineering, the inducibility of genome engineering tools could enable the system to tightly monitor and regulate each component of the relevant metabolic pathways throughout the manufacturing cycle (*145-149*).

6.3 Other future directions for improving the functional versatility of genome engineering technologies

In addition to precision restriction of activity, there is also huge space for future development of genome engineering technologies with regard to the engineering activity itself. Because both technological platforms developed in my research work, TALE and CRISPR/Cas systems, are capable of serving as a general approach for targeting specific loci within the mammalian genomes, there is a large number of possibilities for using these genome-targeting tools for controlling other aspects of the sophisticated genome organization in addition to the alternation of DNA sequence or modulating the expression level of particular genes. This relies on the fact that the TALE DNA binding domain, which contains the N-terminus and the repetitive region, is sufficient for binding to target genomic sites, whereas recent work on CRISPR-Cas system revealed that the nuclease-activity-compromised version of the Cas9 enzyme, i.e., with mutations in either of its two catalytic RuvC and HNH nuclease domains, is still bound to the target DNA substrate when directed by the guide RNA (*29, 121, 150*). Hence, the fusion protein strategies delineated in the introduction part of this report could be applied to both technologies for expanding their power for genome engineering (*1, 2, 5*).

Indeed, while most current efforts focus on increasing the efficiency for the introduction of genetic variants into the genome with either improved nuclease domains or cooperative activities of auxiliary factors or delivery methodologies, a growing number of researchers have started to pay attention to other potential capacities of these systems within a broader definition of genome engineering, e.g. the modulation of transcriptional

activity within mammalian cells, the regulation of a variety of epigenetic markers and in larger scale the chromatin states of the mammalian genomes, and the implementation of a designer integrase/recombinase that allows scarless, higher-order engineering at the whole genome scale with minimal off-target effects (*1, 3, 13, 14, 23, 25, 29, 48, 51, 53, 73, 84, 103, 117, 135, 151-153*).

6.4 Application of genome engineering in disease modeling and the development of human gene therapy for currently untreatable diseases

One of our preeminent and long-term goal for developing efficient and accurate genome engineering technologies is the application of these tools in molecular and cellular medicine. Here I would focus the discussion on two major efforts in this field, creating models for complex disorders and developing gene therapies for various human diseases.

The underlying genetic causes of a variety of human diseases are being investigated through large-scale sequencing or genome-wide association studies with an urgent need for validation of putative causes of disease and determination of effective targets for drug development. Taking the example of neuro-psychiatric diseases, they together forms one of the most devastating and prevalent categories of human disorders in the modern times as measured by the physical and mental burdens inflicted by these diseases at personal, family, and society levels (*154-158*). A major consensus now for the patho-physiology of neuro-psychiatric diseases is that they arise from complex interactions of multiple genetic and epigenetic factors. Significant heritability of several neuro-psychiatric diseases like schizophrenia, major depressive diseases (MDDs), and autism spectrum disorders

(ASDs) indicates the importance of creating disease models based on such information (*159-161*). Notwithstanding years of investigation and the maturity of numerous powerful read-out technologies, we are still at the beginning of exploring this area due to technological limits in the creation of disease models. Now with this new generation of genome engineering tools based on TALEs and CRISPR-Cas systems, it is feasible to quickly build thousands of target-specific nucleases or epigenetic modulators to generate potentially all the exact genomic or epigenomic variants in a model systems so that the biochemical, physiological, and behavior implications of these variants could be thoroughly studied in appropriate contexts and experimental set-ups (*53, 60, 89, 118, 151, 162-165*). Hence, the process of using these technologies to generate large library of model organisms could significantly improve the cost-effectiveness, duration, and precision for disease modeling, and thereby speed up the progress of verifying *bona fide* disease causes, drug targets and screening these targets for new therapeutics (*1, 14, 17, 36, 60*).

More recently, the emerging application of stem cell technology in disease modeling is bringing a new wave of interactions among biotechnolgies for tackling human diseases. We now have the ability to reprogram somatic cells and re-differentiate them into diverse cell types (*166-168*). Given initial evidence that induced pluripotent stem cells (iPSCs) exhibit the hallmarks of disease phenotype at cellular level and the fact that the molecular basis of many neuro-psychiatric diseases are developmental, iPSCs might hold the promise of serving as an ideal existing technology to combine with these genome engineering technology (*168-172*). The iPSCs will enable a fast and universal platform of

testing the effects of disease-relevant variants or confirming their involvement through phenotypic rescue experiments. They ultimately could be combined with aforementioned whole-animal *in vivo* studies to fully realize its potential in therapeutic development.

On the other hand, the more direct way of applying genome engineering technologies for medicine is gene therapy. There have been many concerns over the effectiveness and safety of gene therapy in the past decades, particularly given the issue with viral vector integration and off-target effects (*173-175*). Nonetheless, new progress in the delivery methods, gene editing tools, and above all the elucidation of underlying genetic causes of a great variety of human diseases, led to perhaps one of the most promising and encouraging time for the development and clinical implementation of gene therapy (*172, 176-185*). After many years of stagnation, we started to see the revive of gene therapy studies and clinical trials around the world. One prominent example for using gene therapy in human disorders is the ongoing phase II/III clinical trial on using zinc finger nucleases to target the human CCR5 gene for treatment and prevention of HIV/AIDS, developed and implemented by Sangamo Biosciences (SGM), which have demonstrated promising results in pre-clinical studies and benefited from relatively clear biological rationale for the design of the therapeutic strategy (*172*). The same approach is being utilized in a few other cases for the development of targeted gene therapy (*85, 172*). Most of the efforts from Sangamo are based on the technology zinc finger nucleases (ZFNs) which has similar basic functional manifestation to the more recent TALE and CRISPR/Cas systems, thus representing a general strategy by which genome engineering tools could be used as clinical treatment or the only cure for many of the most difficult-

to-conquer human hereditary disorders that have hardly any effective existing therapies (*151, 176-178*).

6.5 Integration of genome engineering technology and its future potential

Overall, the integration of technologies based on my work from all three parts, along with other existing genome engineering tools and a variety of readout methods, could form a truly transforming technology platform that can improve our understanding of biology and human diseases. Future endeavor built upon these genome-scale high-throughput tools will enable more powerful applications. The impact of these integrations range from the detailed knowledge of fine structures and regulations of the human genome, to the building of reliable disease models, to groundbreaking therapeutics of human disorders, to ultimately improve the quality of life for every member of our society.

Appendix A. Supplementary Information for Chapter 2

Additional methods

Endogenous gene activation assay in mouse cell line

Endogenous gene activation assay: neuro-2a cells were seeded in 24 well plates. 1ug of dTALE plasmid was transfected using Lipofectamine 2000 (Life Technologies). Transfected cells were cultured at 37°C for 48 hours. Cells were harvested and subsequently processed for total RNA extraction using the RNAeasy Plus Mini Kit (Qiagen). cDNA was generated using the qScript cDNA supermix (Quanta Bio) according to the manufacturer's recommended protocol. Oct4 and cMyc mRNA were detected using TaqMan Gene Expression Assays (Life Technologies: Oct4 - Mm03053917_g1, cMyc - Mm00487804_m1) following the manufacturer's protocol using Taqman Advanced MasterMix (Life Technologies).

Design and synthesis strategy for designer TALEs:

Designer TALEs with customized DNA binding domains were constructed using hierarchical ligation as described below.

Step 1: Optimization of DNA sequence for each repeat monomer to minimize repetitiveness of the final product.

Repetitive DNA sequences are difficult to manipulate for a number of reasons, including susceptibility to recombination and difficulty for PCR amplification. To reduce the repetitiveness of designer TALEs, we first optimized the DNA sequence of the four monomers (NI, HD, NN, NG) to minimize homology while preserving the amino acid

sequence. We used the 34aa repeat monomer from the *Xanthomonas sp*. hax3 gene and generated 4 monomers (Supplementary Table 1). The new monomers were synthesized (DNA2.0, Menlo Park, CA) and cloned into individual plasmids to be used as amplification templates.

Step 2: Design of a ligation strategy that utilizes orthogonal sticky ends to specify the position of each monomer in the ligated tandem repeat.

In order to assemble the individual monomers in a specific order, we altered the DNA sequence at the junction between each pair of monomers, similar to the Golden Gate cloning method(*43, 44*). The junction is Gly-Leu and has a total of 24 possible codon pairs (4 codons for Gly and 6 codons for Leu) with 4 variable bases. We initially chose 12 ligation linkers and found that the ability to ligate 12 pieces together into the specified order was very inefficient. Therefore, we tested multi-piece ligation reactions containing 2 to 12 pieces and found that 4-piece ligation was most efficient. This led us to revise our assembly strategy using hierarchical ligation where three 4-mer tandem repeats were assembled first, and the pre-assembled 4-mer tandem repeats were subsequently ligated to form the final 12-mer tandem repeats. To minimize the formation of incorrectly ligated products, we tested a series of ligases and found that T7 ligase gave the highest efficiency for multi-piece ligation. For specifying the order of the 4 monomers, we used the following linkers to construct each 4-mer tandem repeat:

AA : G L G L G L

```
DNA:   5′ ............GGA CTC........GGC CTC........GGA TTA............
3′
       3′ ............CCT GAG........CCG GAG........CCT AAT............
5′
          |---repeat 1---|---repeat 2---|---repeat 3---|---repeat 4---|
```

The 4-mers were used as assembly blocks for constructing the full 12-mer repeat. The completely assembled 12-mer was cloned into the appropriate destination plasmid containing the N- and C-termini of hax3 as well as one 0.5 repeat. The ligated tandem repeat along with the N- and C-termini form a fully functional designer TALE. The sequences for the destination plasmids are listed in Supplementary Sequences.

Step 3: Construction of designer TALEs

We used PCR amplification to generate a set of monomers (HD, NI, NN, NG) for each position. We designed a set of 24 primers (Supplementary Table 2) to attach the right ligation junction onto each monomer. Using type IIs enzymes (e.g. *Bsm*BI and BsaI) we can process the ends of each repeat monomer to expose the sticky-end ligation junctions. An example is:

```
AA:                 L  T  P  E  Q  V  V  .....  C  Q  A  H  G  L
DNA:   5′ cgtctcGACTCACCCCAGAGCAGGTCGTG.......TGCCAAGCGCACGGCCTCAgagacc
       3′ gcagagCTGAGTGGGGTCTCGTCCAGCAC.......ACGGTTCGCGTGCCGGAGTctctgg
           BsmBI Site                                        BsaI Site

                                    |
                                    |   Digest with BsaI
                                    V
```

```
AA:                  L  T  P  E  Q  V  V  .....  C  Q  A  H  G  L
DNA:           5′ ACTCACCCCAGAGCAGGTCGTG.......TGCCAAGCGCACGG
                  3′     TGGGGTCTCGTCCAGCAC.......ACGGTTCGCGTGCCGGAG
```

Step 4: Ligate monomers into specific tandem repeats (Fig. 1b). We constructed 12mer tandem repeats in two steps. 4mer tandem repeats were first assembled in 10ul ligation reactions consisting of 25ng for each monomer. We specifically chose the T7 DNA ligase based on its 1000x higher activity on sticky ends than blunt ends. The correct size ligation product (~440bp) for each 4mer tandem repeat is then purified and PCR amplified. The 4mer PCR products were then processed with the appropriate type IIs enzyme and then ligated again to form 12mer tandem repeats. The correctly ligated 12mer product is then PCR amplified again and processed with Type IIs enzyme for ligation into the backbone plasmid. The final assembled dTALE is verified by sequencing.

Material and procools for building designer TALEs

The original set of primers (Supplementary Table 2) was designed for optimizing the dTALE assembly procedure. Each monomer had two different restriction sites flanking the 5' and 3' ends. We tested a number of ligation conditions by varying the number of pieces being ligated simultaneously and found that 4-piece ligation works most efficiently. To avoid the need for double digests using two different restriction enzymes at different temperatures, we revised the original dTALE construction protocol as well as

primer design (Supplementary Table 3) to streamline the assembly process. A step-by-step protocol for the simplified dTALE construction method is presented below:

1. A library consisting of 48 monomers (4 monomers for each position in the final assembled 12-mer tandem repeat) is generated using PCR. Plasmids containing each type of monomer repeat (monomer sequence listed in Supplementary Table 1) are used as the template for amplification. PCR reactions are set up for each monomer, according to the following table of template and primer pairing (primers shown in Supplementary Table 3).

Primer:	F1/R1	F2/R2	F3/R3	F4/R4	F5/R5	F6/R6	F7/R7	F8/R8	F9/R9	F10/R10	F11/R11	F12/R12
Template:	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI
	F1/R1	F2/R2	F3/R3	F4/R4	F5/R5	F6/R6	F7/R7	F8/R8	F9/R9	F10/R10	F11/R11	F12/R12
	HD	HD	HD	HD	HD	HD	HD	HD	HD	HD	HD	HD
	F1/R1	F2/R2	F3/R3	F4/R4	F5/R5	F6/R6	F7/R7	F8/R8	F9/R9	F10/R10	F11/R11	F12/R12
	NG	NG	NG	NG	NG	NG	NG	NG	NG	NG	NG	NG
	F1/R1	F2/R2	F3/R3	F4/R4	F5/R5	F6/R6	F7/R7	F8/R8	F9/R9	F10/R10	F11/R11	F12/R12
	NN	NN	NN	NG	NN	NN	NN	NN	NN	NN	NN	NN

2. For monomer PCR, high-fidelity polymerase (e.g. Herculase II) (Stratagene) is used to minimize mutation and achieve the highest product yield. Monomers are amplified in 100ml PCR reactions following appropriate protocols of polymerase manufacturers.

3. After completion of the PCR reaction, each monomer is purified using the 96 QIAquick PCR Purification Kit (Qiagen) and the product eluted in 70ml of ddH_2O.

4. Each monomer is digested using *Bsa*I (New England BioLabs) at 37°C for 1 hour in a 100ml reactions as follows:

70ml	purified PCR Product
5ml	BsaI (50 units)
5ml	10X Buffer #4
1ml	100X BSA
19ml	ddH_2O

5. After digestion, digested monomers are purified using the 96 QIAquick PCR Purification Kit (Qiagen) and eluted in 70ml of ddH_2O.

6. The concentration of each monomer is adjusted to 25ng/ml for monomers 1, 4, 5, 8, 9, 12; and 20ng/ml for monomers 2,3,6,7,10,11.

7. For each dTALE to be assembled, individual 4-mer tandem repeats are first constructed by simultaneously ligating four repeat monomers together at equal molar ratio (25ng for monomers 1 and 4, 20ng for monomers 2 and 3 in 10ul total ligation mix, using 300units of T7 ligase from Enzymatics and 10X ligation buffer). The ligation is incubated at room temperature for 30 minutes.

8. 5ml of the 4-mer ligation reactions are run on a 2% E-Gel EX (Invitrogen) and the correct size products are amplified by gel-stab PCR(*186*). Specifically, a 10 μL

pipette tip is used to puncture the gel at the location of the desired product. The stab is mixed up and down in 10 μL of water, and the water is heated to 65 °C for 2 min. 2.5 μL of the gel-isolated product diluted in water is then amplified in a 50 μL PCR reaction using Herculase II polymerase (Stratagene).

9. The amplified 4-mer tandem repeats are purified using the QIAquick PCR Purification kit and eluted in 40ml of ddH_2O.

10. Purified 4-mer tandem repeats as well as the appropriate dTALE backbone are digested using *Bsm*BI at 55°C for 1 hour.

40ul	purified PCR product		500ng	dTALE backbone vector	
5ul	10X Buffer #3	and	5ul	10X Buffer #3	
5ul	BsmBI (50 units)		5ul	*Bsm*BI (50 units)	
	ddH_2O)			(bring volume to 50ul with	

11. Purify digested 4-mer tandem repeats using QIAquick PCR Purification Kit (Qiagen). Gel purify the digested backbone.

12. Fully assembled dTALEs are generated by simultaneously ligating the three 4-mer tandem repeats with the backbone vector at equal molar ratio (1ng for each 4-mer tandem repeat and 28ng for backbone vector; in a 10ul ligation reaction using

1500U T7 ligase from Enzymatics). A negative control reaction should be set up with 28ng of the backbone vector alone. All ligation reactions are incubated in a thermal cycler using the following parameters: 37^{o}C for 1 min followed by 25^{o}C for 5min, for 30 cycles.

13. 2ul of each dTALE ligation reaction is transformed into XL-10 Gold chemically competent cells (Stratagene).

14. Plasmid DNA for assembled dTALE are prepared from ligation transformants and analyzed via restriction digest and DNA sequencing. Transformation of the negative control ligation should not yield any transformants.

Supplementary Table 1 | List of repeat monomer sequences. Forward and reverse priming sites are highlighted in blue and red respectively.

NI

L T P E Q V V A I A S **N** **I** G G K Q A L

CTG**ACCCCAGAGCAGGTCGTG**GCAATCGCCTCCAACATTGGCGGGAAACAGGCACTC

E T V Q R L L P V L C Q A H G

GAGACTGTCCAGCGCCTGCTTCCCGTGCT**GTGCCAAGCGCACGG**A

HD

L T P E Q V V A I A S **H** **D** G G K Q A L

TTG**ACCCCAGAGCAGGTCGTG**GCGATCGCAAGCCACGACGGAGGAAAGCAAGCCTTG

E T V Q R L L P V L C Q A H G

GAAACAGTACAGAGGCTGTTGCCTGTGCT**GTGCCAAGCGCACGG**G

NN

L T P E Q V V A I A S **N** **N** G G K Q A L

CTT**ACCCCAGAGCAGGTCGTG**GCAATCGCGAGCAATAACGGCGGAAAACAGGCTTTG

E T V Q R L L P V L C Q A H G

GAAACGGTGCAGAGGCTCCTTCCAGTGCT**GTGCCAAGCGCACGG**G

NG

L T P E Q V V A I A S **N** **G** G G K Q A L

CTG**ACCCCAGAGCAGGTCGTG**GCCATTGCCTCGAATGGAGGGGGCAAACAGGCGTTG

E T V Q R L L P V L C Q A H G

GAAACCGTACAACGATTGCTGCCGGTGCT**GTGCCAAGCGCACGG**C

Supplementary Table 2 | Primers used for the amplification and assembly of artificial TALEs reported in this manuscript. Unique linkers used to specify the ligation ordering are highlighted in blue; sequences written from 5' to 3'. *Bsm*BI sites are highlighted in yellow and *Bsa*I sites are highlighted in gray.

F1	AGATGCCGTCCTAGCGcgtctcCTGACCCCAGAGCAGGTCGTGG
F2	AGATGCCGTCCTAGCGcgtctcGACTCACCCCAGAGCAGGTCGTG
F3	AGATGCCGTCCTAGCGcgtctcGCCTCACCCCAGAGCAGGTCGTG
F4	AGATGCCGTCCTAGCGcgtctcGATTAACCCCAGAGCAGGTCGTG
F5	AGATGCCGTCCTAGCGcgtctcGCTTAACCCCAGAGCAGGTCGTG
F6	AGATGCCGTCCTAGCGcgtctcGACTTACCCCAGAGCAGGTCGTG
F7	AGATGCCGTCCTAGCGcgtctcGCCTTACCCCAGAGCAGGTCGTG
F8	AGATGCCGTCCTAGCGcgtctcGACTAACCCCAGAGCAGGTCGTG
F9	AGATGCCGTCCTAGCGcgtctcGGCTCACCCCAGAGCAGGTCGTG
F10	AGATGCCGTCCTAGCGcgtctcGGCTAACCCCAGAGCAGGTCGTG
F11	AGATGCCGTCCTAGCGcgtctcGCCTAACCCCAGAGCAGGTCGTG
F12	AGATGCCGTCCTAGCGcgtctcGGTTAACCCCAGAGCAGGTCGTG
R1	GTATCTTTCCTGTGCCCAggtctcTGAGTCCGTGCGCTTGGCAC
R2	GTATCTTTCCTGTGCCCAggtctcTGAGGCCGTGCGCTTGGCAC
R3	GTATCTTTCCTGTGCCCAggtctcTTAATCCGTGCGCTTGGCAC
R4	GTATCTTTCCTGTGCCCAggtctcTTAAGCCGTGCGCTTGGCAC
R5	GTATCTTTCCTGTGCCCAggtctcTAAGTCCGTGCGCTTGGCAC
R6	GTATCTTTCCTGTGCCCAggtctcTAAGGCCGTGCGCTTGGCAC
R7	GTATCTTTCCTGTGCCCAggtctcTTAGTCCGTGCGCTTGGCAC

(Supplementary Table 2, continued)

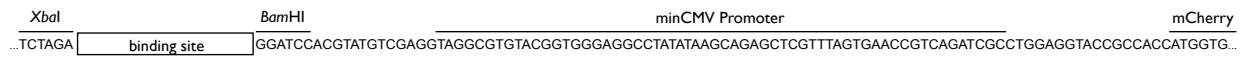
R8	GTATCTTTCCTGTGCCCAggtctcTGAGCCCGTGCGCTTGGCAC
R9	GTATCTTTCCTGTGCCCAggtctcTTAGCCCGTGCGCTTGGCAC
R10	GTATCTTTCCTGTGCCCAggtctcTTAGGCCGTGCGCTTGGCAC
R11	GTATCTTTCCTGTGCCCAggtctcTTAACCCGTGCGCTTGGCAC
R12	GTATCTTTCCTGTGCCCAggtctcTTAAACCGTGCGCTTGGCAC
F-assem	ATATAGATGCCGTCCTAGCGC
R-assem	AAGTATCTTTCCTGTGCCCAG

Supplementary Table 3 | Primers used in the simplified step-by-step dTALE construction method. Unique linkers used to specify the ligation ordering are highlighted in blue; sequences written from 5' to 3'. *Bsm*BI sites are highlighted in yellow and *Bsa*I sites are highlighted in gray.

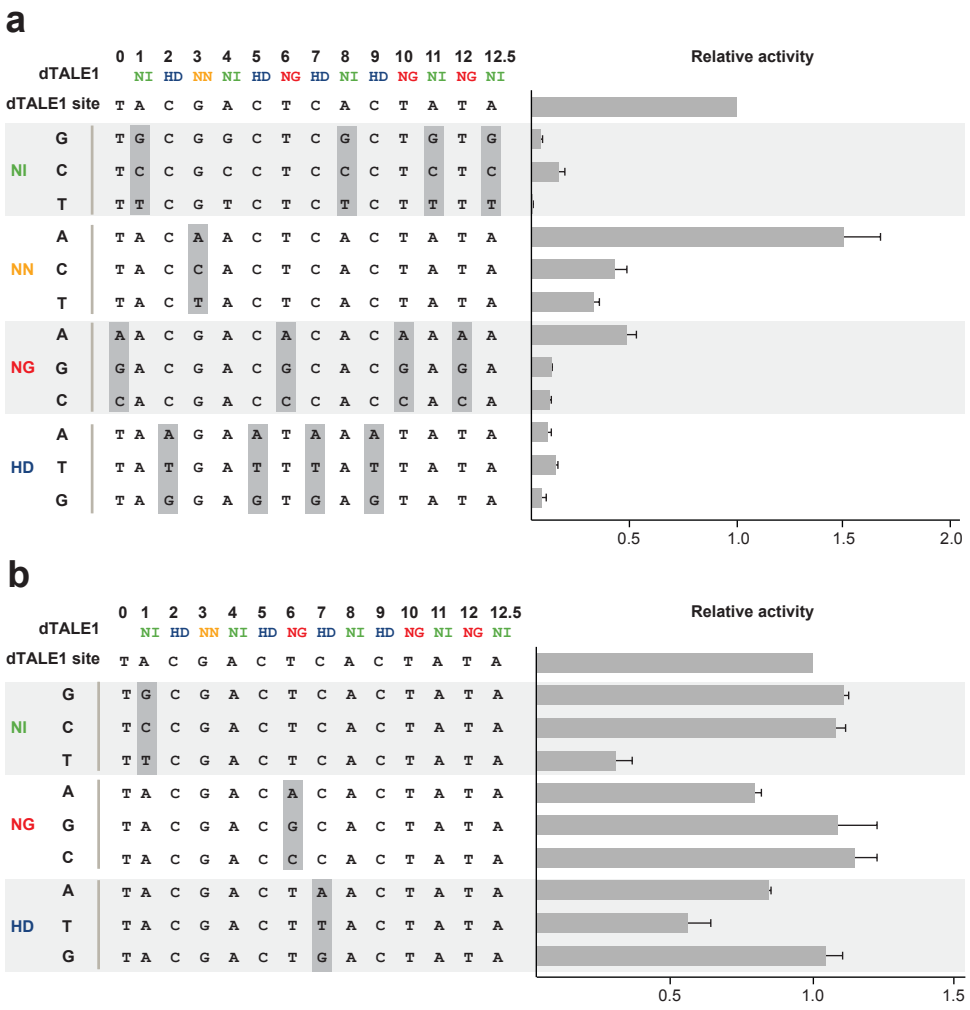
F1	ATATAGATGCCGTCCTAGCGcgtctcCTGACCCCAGAGCAGGTCGTGG
F2	TGCTCTTTATTCGTTGCGTCggtctcGACTCACCCCAGAGCAGGTCGTG
F3	TGCTCTTTATTCGTTGCGTCggtctcGCCTCACCCCAGAGCAGGTCGTG
F4	TGCTCTTTATTCGTTGCGTCggtctcGATTAACCCCAGAGCAGGTCGTG
F5	ATATAGATGCCGTCCTAGCGcgtctcGCTTAACCCCAGAGCAGGTCGTG
F6	TGCTCTTTATTCGTTGCGTCggtctcGACTCACCCCAGAGCAGGTCGTG
F7	TGCTCTTTATTCGTTGCGTCggtctcGCCTCACCCCAGAGCAGGTCGTG
F8	TGCTCTTTATTCGTTGCGTCggtctcGATTAACCCCAGAGCAGGTCGTG
F9	ATATAGATGCCGTCCTAGCGcgtctcGGCTCACCCCAGAGCAGGTCGTG
F10	TGCTCTTTATTCGTTGCGTCggtctcGACTCACCCCAGAGCAGGTCGTG
F11	TGCTCTTTATTCGTTGCGTCggtctcGCCTCACCCCAGAGCAGGTCGTG
F12	TGCTCTTTATTCGTTGCGTCggtctcGATTAACCCCAGAGCAGGTCGTG
R1	TCTTATCGGTGCTTCGTTCTggtctcTGAGTCCGTGCGCTTGGCAC
R2	TCTTATCGGTGCTTCGTTCTggtctcTGAGGCCGTGCGCTTGGCAC
R3	TCTTATCGGTGCTTCGTTCTggtctcTTAATCCGTGCGCTTGGCAC
R4	AAGTATCTTTCCTGTGCCCAcgtctcTTAAGCCGTGCGCTTGGCAC
R5	TCTTATCGGTGCTTCGTTCTggtctcTGAGTCCGTGCGCTTGGCAC
R6	TCTTATCGGTGCTTCGTTCTggtctcTGAGGCCGTGCGCTTGGCAC

(Supplementary Table 3, continued)

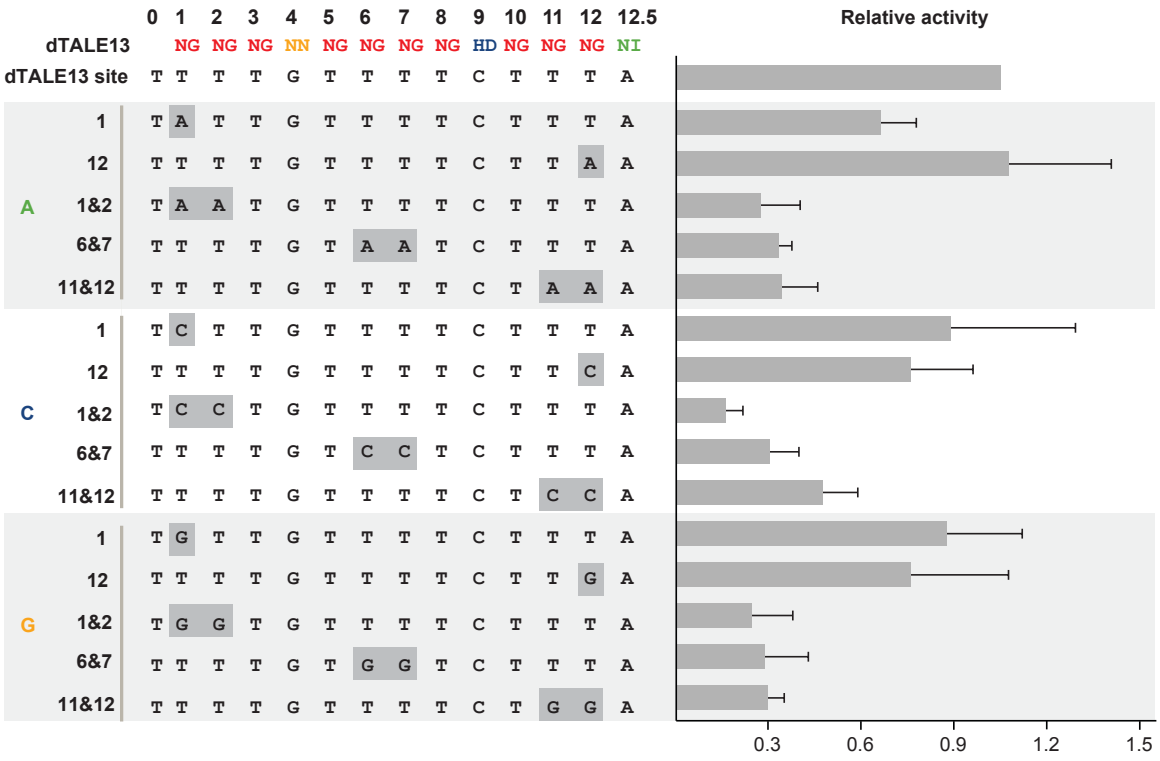
R7	TCTTATCGGTGCTTCGTTCTggtctcTTAATCCGTGCGCTTGGCAC
R8	AAGTATCTTTCCTGTGCCCAcgtctcTGAGCCCGTGCGCTTGGCAC
R9	TCTTATCGGTGCTTCGTTCTggtctcTGAGTCCGTGCGCTTGGCAC
R10	TCTTATCGGTGCTTCGTTCTggtctcTGAGGCCGTGCGCTTGGCAC
R11	TCTTATCGGTGCTTCGTTCTggtctcTTAATCCGTGCGCTTGGCAC
R12	AAGTATCTTTCCTGTGCCCAcgtctcTGAGTCCGTGCGCTTGGCAC
F-assem	ATATAGATGCCGTCCTAGCG
R-assem	AAGTATCTTTCCTGTGCCCA



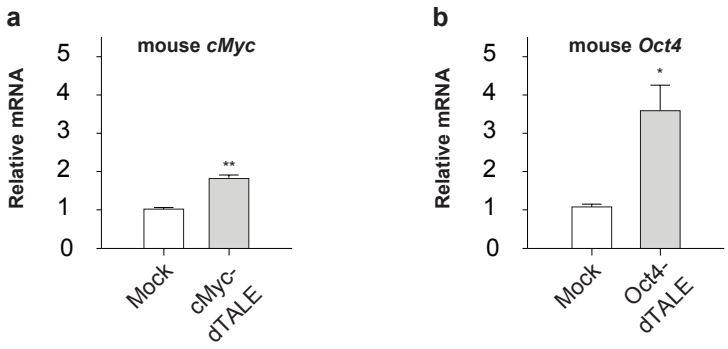
Supplementary Figure 1 | Design of mCherry reporter plasmid. Target binding site of a dTALE was cloned into the mCherry reporter plasmid between the XbaI and BamHI restriction sites. Hence, the dTALE binding site is placed -96bp upstream of the transcription start site of a full-length mCherry gene, with a minimal CMV promoter in the middle.



Supplementary Figure 2 | Test of the DNA binding specificity of dTALE using reporters with varying numbers of mismatches. a, Each diresidue in dTALE1 was tested against its non-preferred DNA bases to determine the binding specificity for each diresidue. b, Single base pair mismatches were used to test the binding specificity of dTALE1. The relative activity of dTALE1 for each mutant reporter compared to the intended reporter is shown on the right. All error bars indicate s.e.m, n=3. The fold induction was determined via flow cytometry analysis of mCherry expression in transfected 293FT cells, and calculated as the ratio of the total mCherry fluorescence intensity of cells transfected with and without the specified dTALE.



Supplementary Figure 3 | Tests of the DNA binding specificity of dTALE to mismatched sequences. DNA binding specificity of dTALE was tested using dTALE13 and a series of reporters bearing systematically designed mutations in the binding site for dTALE13. The design of the reporter series is shown on the left, with three different groups: A, C, and G. Within each group, the base T at one or two positions of the dTALE-binding site (designated by numbers to the left of each target sequence) were altered to A, C, or G (the mutated bases are highlighted). The relative activity of dTALE13 for each mutant reporter compared to the original reporter is shown on the right. All error bars indicate s.e.m, n=3. The fold induction was determined via flow cytometry analysis of mCherry expression in transfected 293FT cells, and calculated as the ratio of the total mCherry fluorescence intensity of cells transfected with and without the specified dTALE.



Supplementary Figure 4 | Activation of endogenous pluripotency factors from the mouse genome by designer TALEs. a, b, mRNA levels of *cMyc* and *Oct4* in mouse neuro-2a cells transfected with mock, cMyc-dTALE and Oct4-dTALE. Bars represent the levels of *cMyc* and *Oct4* mRNA in the transfected cell as determined via quantitative RT-PCR. Mock consists of cells receiving the transfection vehicle. All error bars indicate s.e.m.; n=3. * $p < 0.5$, ** $p < 0.05$, un-paired student's t test.

Supplementary sequences

Type IIs sites are colored in blue. NLS is colored in red. 2A-GFP is colored in green. The variable diresidue is highlighted yellow.

>dTALE-Backbone(NI in 0.5 repeat)-NLS-VP64-2A-EGFP

ATGTCGCGGACCCGGCTCCCTTCCCCACCCGCACCCAGCCCAGCGTTTTCGGCCGACTCGTTCTCAGACCTGCTTA
GGCAGTTCGACCCCTCACTGTTTAACACATCGTTGTTCGACTCCCTTCCTCCGTTTGGGGCGCACCATACGGAGGCG
GCCACCGGGGAGTGGGATGAGGTGCAGTCGGGATTGAGAGCTGCGGATGCACCACCCCCAACCATGCGGGTGGCC
GTCACCGCTGCCCGACCGCCGAGGGCGAAGCCCGCACCAAGGCGGAGGGCAGCGCAACCGTCCGACGCAAGCCC
CGCAGCGCAAGTAGATTTGAGAACTTTGGGATATTCACAGCAGCAGCAGGAAAAGATCAAGCCCAAAGTGAGGTCGA
CAGTCGCGCAGCATCACGAAGCGCTGGTGGGTCATGGGTTTACACATGCCCACATCGTAGCCTTGTCGCAGCACCC
TGCAGCCCTTGGCACGGTCGCCGTCAAGTACCAGGACATGATTGCGGCGTTGCCGGAAGCCACACATGAGGCGATC
GTCGGTGTGGGGAAACAGTGGAGCGGAGCCCGAGCGCTTGAGGCCCTGTTGACGGTCGCGGGAGAGCTGAGAGG
GCCTCCCCTTCAGCTGGACACGGGCCAGTTGCTGAAGATCGCGAAGCGGGGAGGAGTCACGGCGGTCGAGGCGGT
GCACGCGTGGCGCAATGCGCTCACGGGAGCACCCCTCAACCTGACCgagacgGTACATGAAACGCATGGCACGGcgtct
cAACTCACGCCTGAGCAGGTAGTGGCTATTGCATCCAATATCGGGGGCAGACCCGCACTGGAGTCAATCGTGGCCC
AGCTTTCGAGGCCGGACCCCGCGCTGGCCGCACTCACTAATGATCATCTTGTAGCGCTGGCCTGCCTCGGCGGACG
ACCCGCCTTGGATGCGGTGAAGAAGGGGCTCCCGCACGCGCCTGCATTGATTAAGCGGACCAACAGAAGGATTCCC
GAGAGGACATCACATCGAGTGGCAGATCACGCGCAAGTGGTCCGCGTGCTCGGATTCTTCCAGTGTCACTCCCACC
CCGCACAAGCGTTCGATGACGCCATGACTCAATTTGGTATGTCGAGACACGGACTGCTGCAGCTCTTTCGTAGAGTC
GGTGTCACAGAACTCGAGGCCCGCTCGGGCACACTGCCTCCCGCCTCCCAGCGGTGGGACAGGATTCTCCAAGCG
AGCGGTATGAAACGCGCGAAGCCTTCACCTACGTCAACTCAGACACCTGACCAGGCGAGCCTTCATGCGTTCGCAG
ACTCGCTGGAGAGGGATTTGGACGCGCCCTCGCCCATGCATGAAGGGGACCAAACTCGCGCGTCAgctagccccaagaa
gaagagaaaggtggaggccagcggttccGGACGGGCTGACGCATTGGACGATTTTGATCTGGATATGCTGGGAAGTGACGCCCT
CGATGATTTTGACCTTGACATGCTTGGTTCGGATGCCCTTGATGACTTTGACCTCGACATGCTCGGCAGTGACGCCC
TTGATGATTTCGACCTGGACATGCTGATTAACtctagaggcagtggagagggcagaggaagtctgctaacatgcggtgacgtcgaggagaatcctg
gcccaGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGG
CCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACC
ACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACC
CCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTC
AAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTG
AAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCT
ATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGT
GCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCT
GAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCC
GCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAA

>dTALE-Backbone(NG in 0.5 repeat)-NLS-VP64-2A-EGFP

ATGTCGCGGACCCGGCTCCCTTCCCCACCCGCACCCAGCCCAGCGTTTTCGGCCGACTCGTTCTCAGACCTGCTTA
GGCAGTTCGACCCCTCACTGTTTAACACATCGTTGTTCGACTCCCTTCCTCCGTTTGGGGCGCACCATACGGAGGCG
GCCACCGGGGAGTGGGATGAGGTGCAGTCGGGATTGAGAGCTGCGGATGCACCACCCCCAACCATGCGGGTGGCC
GTCACCGCTGCCCGACCGCCGAGGGCGAAGCCCGCACCAAGGCGGAGGGCAGCGCAACCGTCCGACGCAAGCCC
CGCAGCGCAAGTAGATTTGAGAACTTTGGGATATTCACAGCAGCAGCAGGAAAAGATCAAGCCCAAAGTGAGGTCGA
CAGTCGCGCAGCATCACGAAGCGCTGGTGGGTCATGGGTTTACACATGCCCACATCGTAGCCTTGTCGCAGCACCC
TGCAGCCCTTGGCACGGTCGCCGTCAAGTACCAGGACATGATTGCGGCGTTGCCGGAAGCCACACATGAGGCGATC
GTCGGTGTGGGGAAACAGTGGAGCGGAGCCCGAGCGCTTGAGGCCCTGTTGACGGTCGCGGGAGAGCTGAGAGG
GCCTCCCCTTCAGCTGGACACGGGCCAGTTGCTGAAGATCGCGAAGCGGGGAGGAGTCACGGCGGTCGAGGCGGT
GCACGCGTGGCGCAATGCGCTCACGGGAGCACCCCTCAACCTGACCgagacgGTACATGAAACGCATGGCACGGcgtct
cAACTCACGCCTGAGCAGGTAGTGGCTATTGCATCCAATGGCGGGGGCAGACCCGCACTGGAGTCAATCGTGGCCC
AGCTTTCGAGGCCGGACCCCGCGCTGGCCGCACTCACTAATGATCATCTTGTAGCGCTGGCCTGCCTCGGCGGACG
ACCCGCCTTGGATGCGGTGAAGAAGGGGCTCCCGCACGCGCCTGCATTGATTAAGCGGACCAACAGAAGGATTCCC
GAGAGGACATCACATCGAGTGGCAGATCACGCGCAAGTGGTCCGCGTGCTCGGATTCTTCCAGTGTCACTCCCACC
CCGCACAAGCGTTCGATGACGCCATGACTCAATTTGGTATGTCGAGACACGGACTGCTGCAGCTCTTTCGTAGAGTC
GGTGTCACAGAACTCGAGGCCCGCTCGGGCACACTGCCTCCCGCCTCCCAGCGGTGGGACAGGATTCTCCAAGCG
AGCGGTATGAAACGCGCGAAGCCTTCACCTACGTCAACTCAGACACCTGACCAGGCGAGCCTTCATGCGTTCGCAG
ACTCGCTGGAGAGGGATTTGGACGCGCCCTCGCCCATGCATGAAGGGGACCAAACTCGCGCGTCAgctagccccaagaa
gaagagaaaggtggaggccagcggttccGGACGGGCTGACGCATTGGACGATTTTGATCTGGATATGCTGGGAAGTGACGCCCT
CGATGATTTTGACCTTGACATGCTTGGTTCGGATGCCCTTGATGACTTTGACCTCGACATGCTCGGCAGTGACGCCC
TTGATGATTTCGACCTGGACATGCTGATTAACtctagaggcagtggagagggcagaggaagtctgctaacatgcggtgacgtcgaggagaatcctg
gcccaGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGG
CCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACC
ACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACC
CCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTC
AAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTG
AAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCT
ATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGT
GCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCT
GAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCC
GCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAA

>dTALE-Backbone(HD in 0.5 repeat)-NLS-VP64-2A-EGFP

ATGTCGCGGACCCGGCTCCCTTCCCCACCCGCACCCAGCCCAGCGTTTTCGGCCGACTCGTTCTCAGACCTGCTTA
GGCAGTTCGACCCCTCACTGTTTAACACATCGTTGTTCGACTCCCTTCCTCCGTTTGGGGCGCACCATACGGAGGCG
GCCACCGGGGAGTGGGATGAGGTGCAGTCGGGATTGAGAGCTGCGGATGCACCACCCCCAACCATGCGGGTGGCC
GTCACCGCTGCCCGACCGCCGAGGGCGAAGCCCGCACCAAGGCGGAGGGCAGCGCAACCGTCCGACGCAAGCCC
CGCAGCGCAAGTAGATTTGAGAACTTTGGGATATTCACAGCAGCAGCAGGAAAAGATCAAGCCCAAAGTGAGGTCGA
CAGTCGCGCAGCATCACGAAGCGCTGGTGGGTCATGGGTTTACACATGCCCACATCGTAGCCTTGTCGCAGCACCC
TGCAGCCCTTGGCACGGTCGCCGTCAAGTACCAGGACATGATTGCGGCGTTGCCGGAAGCCACACATGAGGCGATC

GTCGGTGTGGGGAAACAGTGGAGCGGAGCCCGAGCGCTTGAGGCCCTGTTGACGGTCGCGGGAGAGCTGAGAGG
GCCTCCCCTTCAGCTGGACACGGGCCAGTTGCTGAAGATCGCGAAGCGGGGAGGAGTCACGGCGGTCGAGGCGGT
GCACGCGTGGCGCAATGCGCTCACGGGAGCACCCCTCAACCTGACCgagacgGTACATGAAACGCATGGCACGGcgtct
cAACTCACGCCTGAGCAGGTAGTGGCTATTGCATCCCATGACGGGGGGCAGACCCGCACTGGAGTCAATCGTGGCCC
AGCTTTCGAGGCCGGACCCCGCGCTGGCCGCACTCACTAATGATCATCTTGTAGCGCTGGCCTGCCTCGGCGGACG
ACCCGCCTTGGATGCGGTGAAGAAGGGGCTCCCGCACGCGCCTGCATTGATTAAGCGGACCAACAGAAGGATTCCC
GAGAGGACATCACATCGAGTGGCAGATCACGCGCAAGTGGTCCGCGTGCTCGGATTCTTCCAGTGTCACTCCCACC
CCGCACAAGCGTTCGATGACGCCATGACTCAATTTGGTATGTCGAGACACGGACTGCTGCAGCTCTTTCGTAGAGTC
GGTGTCACAGAACTCGAGGCCCGCTCGGGCACACTGCCTCCCGCCTCCCAGCGGTGGGACAGGATTCTCCAAGCG
AGCGGTATGAAACGCGCGAAGCCTTCACCTACGTCAACTCAGACACCTGACCAGGCGAGCCTTCATGCGTTCGCAG
ACTCGCTGGAGAGGGATTTGGACGCGCCCTCGCCCATGCATGAAGGGGACCAAACTCGCGCGTCAgctagccccaagaa
gaagagaaaggtggaggccagcggttccGGACGGGCTGACGCATTGGACGATTTTGATCTGGATATGCTGGGAAGTGACGCCCT
CGATGATTTTGACCTTGACATGCTTGGTTCGGATGCCCTTGATGACTTTGACCTCGACATGCTCGGCAGTGACGCCC
TTGATGATTTCGACCTGGACATGCTGATTAACtctagaggcagtggagagggcagaggaagtctgctaacatgcggtgacgtcgaggagaatcctg
gcccaGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGG
CCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACC
ACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACC
CCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTC
AAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTG
AAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCT
ATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGT
GCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCT
GAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCC
GCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAA

>dTALE-Backbone(NN in 0.5 repeat)-NLS-VP64-2A-EGFP

ATGTCGCGGACCCGGCTCCCTTCCCCACCCGCACCCAGCCCAGCGTTTTCGGCCGACTCGTTCTCAGACCTGCTTA
GGCAGTTCGACCCCTCACTGTTTAACACATCGTTGTTCGACTCCCTTCCTCCGTTTGGGGCGCACCATACGGAGGCG
GCCACCGGGGAGTGGGATGAGGTGCAGTCGGGATTGAGAGCTGCGGATGCACCACCCCCAACCATGCGGGTGGCC
GTCACCGCTGCCCGACCGCCGAGGGCGAAGCCCGCACCAAGGCGGAGGGCAGCGCAACCGTCCGACGCAAGCCC
CGCAGCGCAAGTAGATTTGAGAACTTTGGGATATTCACAGCAGCAGCAGGAAAAGATCAAGCCCAAAGTGAGGTCGA
CAGTCGCGCAGCATCACGAAGCGCTGGTGGGTCATGGGTTTACACATGCCCACATCGTAGCCTTGTCGCAGCACCC
TGCAGCCCTTGGCACGGTCGCCGTCAAGTACCAGGACATGATTGCGGCGTTGCCGGAAGCCACACATGAGGCGATC
GTCGGTGTGGGGAAACAGTGGAGCGGAGCCCGAGCGCTTGAGGCCCTGTTGACGGTCGCGGGAGAGCTGAGAGG
GCCTCCCCTTCAGCTGGACACGGGCCAGTTGCTGAAGATCGCGAAGCGGGGAGGAGTCACGGCGGTCGAGGCGGT
GCACGCGTGGCGCAATGCGCTCACGGGAGCACCCCTCAACCTGACCgagacgGTACATGAAACGCATGGCACGGcgtct
cAACTCACGCCTGAGCAGGTAGTGGCTATTGCATCCAATAACGGGGGCAGACCCGCACTGGAGTCAATCGTGGCCC
AGCTTTCGAGGCCGGACCCCGCGCTGGCCGCACTCACTAATGATCATCTTGTAGCGCTGGCCTGCCTCGGCGGACG
ACCCGCCTTGGATGCGGTGAAGAAGGGGCTCCCGCACGCGCCTGCATTGATTAAGCGGACCAACAGAAGGATTCCC
GAGAGGACATCACATCGAGTGGCAGATCACGCGCAAGTGGTCCGCGTGCTCGGATTCTTCCAGTGTCACTCCCACC

CCGCACAAGCGTTCGATGACGCCATGACTCAATTTGGTATGTCGAGACACGGACTGCTGCAGCTCTTTCGTAGAGTC
GGTGTCACAGAACTCGAGGCCCGCTCGGGCACACTGCCTCCCGCCTCCCAGCGGTGGGACAGGATTCTCCAAGCG
AGCGGTATGAAACGCGCGAAGCCTTCACCTACGTCAACTCAGACACCTGACCAGGCGAGCCTTCATGCGTTCGCAG
ACTCGCTGGAGAGGGATTTGGACGCGCCCTCGCCCATGCATGAAGGGGACCAAACTCGCGCGTCAgctagccccaagaa
gaagagaaaggtggaggccagcggttccGGACGGGCTGACGCATTGGACGATTTTGATCTGGATATGCTGGGAAGTGACGCCCT
CGATGATTTTGACCTTGACATGCTTGGTTCGGATGCCCTTGATGACTTTGACCTCGACATGCTCGGCAGTGACGCCC
TTGATGATTTCGACCTGGACATGCTGATTAACtctagaggcagtggagagggcagaggaagtctgctaacatgcggtgacgtcgaggagaatcctg
gcccaGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGG
CCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACC
ACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACC
CCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTC
AAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTG
AAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCT
ATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGT
GCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCT
GAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCC
GCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAA

Appendix B. Supplementary Information for Chapter 5

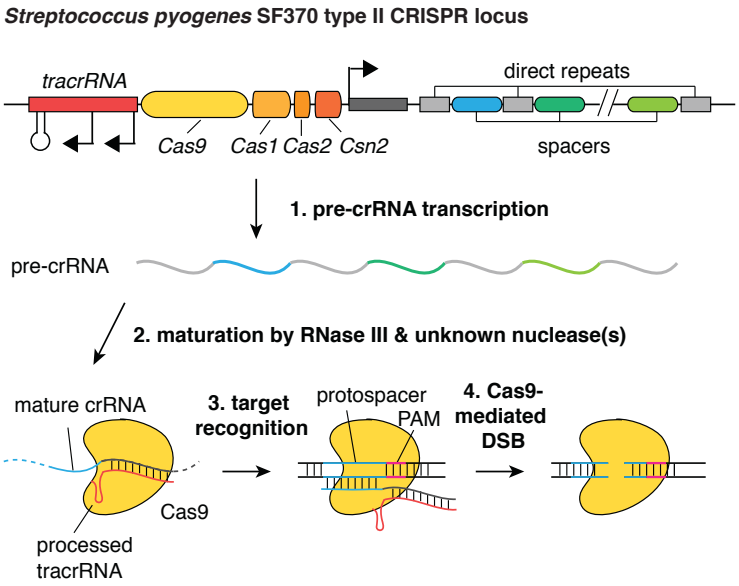


Fig. S1. Schematic of the type II CRISPR-mediated DNA double-strand break. The type II CRISPR locus from *Streptococcus pyogenes* SF370 contains a cluster of four genes, *Cas9*, *Cas1*, *Cas2*, and *Csn1*, as well as two non-coding RNA elements, tracrRNA and a characteristic array of repetitive sequences (direct repeats) interspaced by short stretches of non-repetitive sequences (spacers, 30bp each) (*31, 32, 122, 123, 187, 188*). Each spacer is typically derived from foreign genetic material (protospacer), and directs the specificity of CRISPR-mediated nucleic acid cleavage. In the target nucleic acid, each protospacer is associated with a protospacer adjacent motif (PAM) whose recognition is specific to individual CRISPR systems (*127, 128*). The Type II CRISPR system carries out targeted DNA double-strand break (DSB) in sequential steps (*30, 120, 121, 125, 126*). First, the pre-crRNA array and tracrRNA are transcribed from the CRISPR locus. Second, tracrRNA hybridizes to the direct repeats of pre-crRNA and associates with Cas9 as a duplex, which mediates the processing of the pre-crRNA into mature crRNAs containing individual, truncated spacer sequences**.** Third, the mature crRNA:tracrRNA duplex directs Cas9 to the DNA target consisting of the protospacer and the requisite PAM via heteroduplex formation between the spacer region of the crRNA and the

(Fig. S1, continued) protospacer DNA. Finally, Cas9 mediates cleavage of target DNA upstream of PAM to create a DSB within the protospacer.

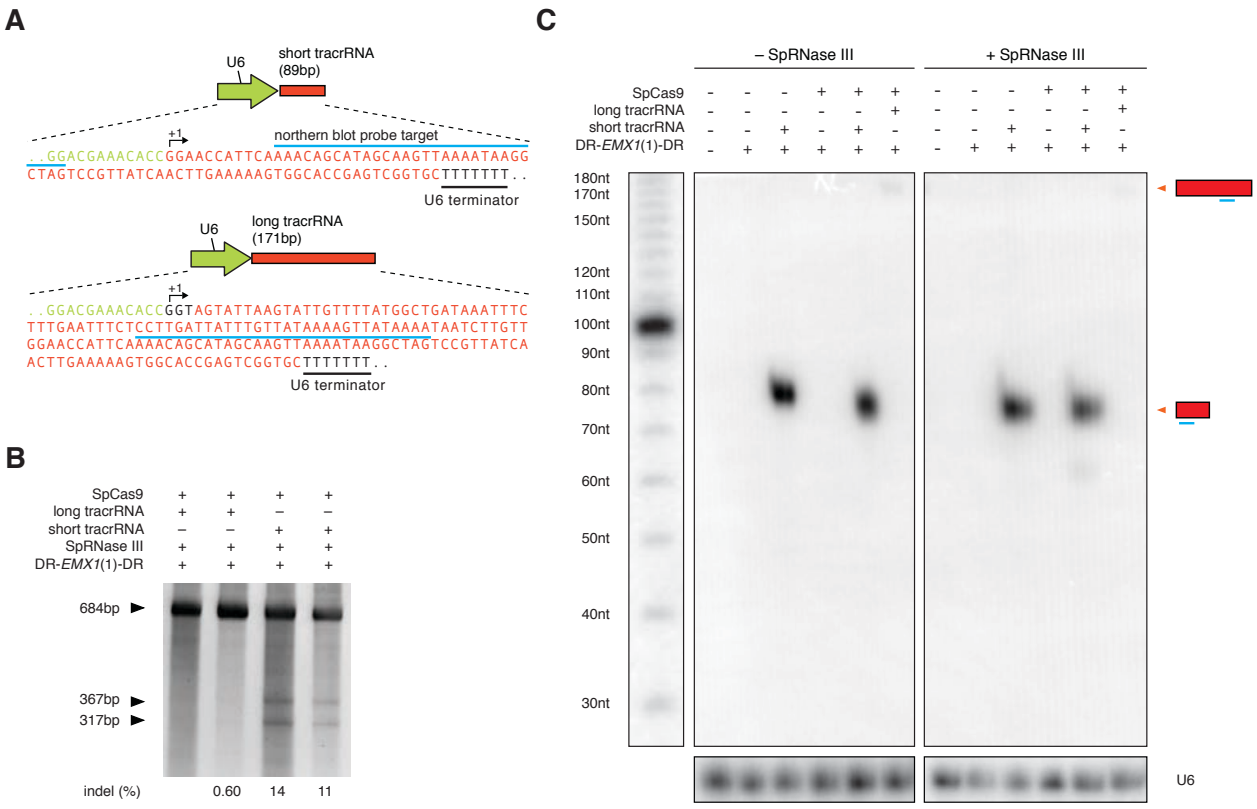


Fig S2. Comparison of different tracrRNA transcripts for Cas9-mediated gene targeting. (A) Schematic showing the design and sequences of two tracrRNA transcripts tested (short and long). Each transcript is driven by a U6 promoter. Transcription start site is marked as +1 and transcription terminator is as indicated. Blue line indicates the region whose reverse-complement sequence is used to generate northern blot probes for tracrRNA detection. (**B**) SURVEYOR assay comparing the efficiency of SpCas9-mediated cleavage of the *EMX1* locus. Two biological replicas are shown for each tracrRNA transcript. (**C**) Northern blot analysis of total RNA extracted from 293FT cells transfected with U6 expression constructs carrying long or short tracrRNA, as well as SpCas9 and DR-*EMX1*(1)-DR. Left and right panels are from 293FT cells transfected without or with SpRNase III respectively. U6 indicate loading control blotted with a probe targeting human U6 snRNA. Transfection of the short tracrRNA expression

(Fig. S2, continued) construct led to abundant levels of the processed form of tracrRNA (~75bp) (*124*). Very low amounts of long tracrRNA are detected on the northern blot. As a result of these experiments, we chose to use short tracrRNA for application in mammalian cells.

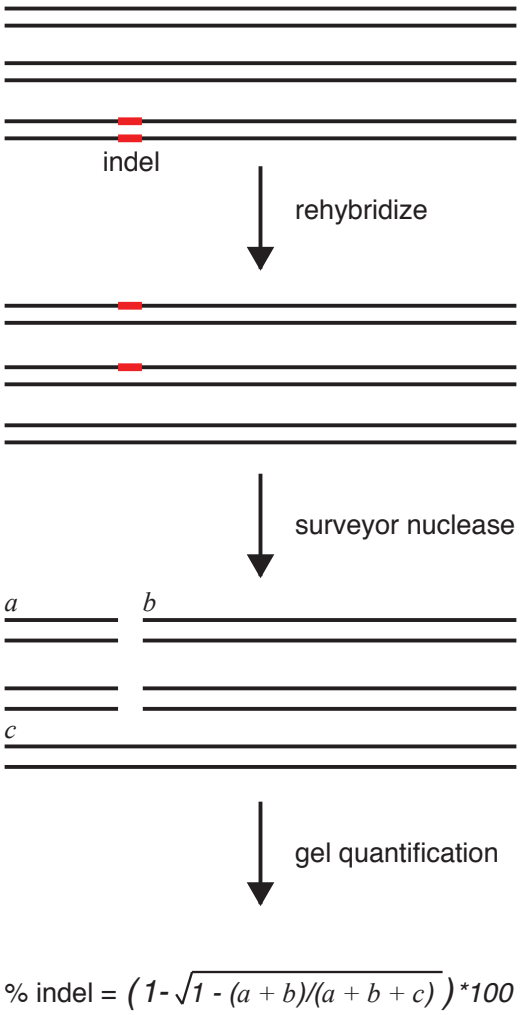


Fig. S3. SURVEYOR assay for detection of double strand break-induced micro insertions and deletions (*93*). Schematic of the SURVEYOR assay used to determine Cas9-mediated cleavage efficiency. First, genomic PCR (gPCR) is used to amplify the Cas9 target region from a heterogeneous population of modified and unmodified cells, and the gPCR products are reannealed slowly to generate heteroduplexes. The reannealed heteroduplexes are cleaved by SURVEYOR nuclease, whereas homoduplexes are left intact. Cas9-mediated cleavage efficiency (% indel) is calculated based on the fraction of cleaved DNA.

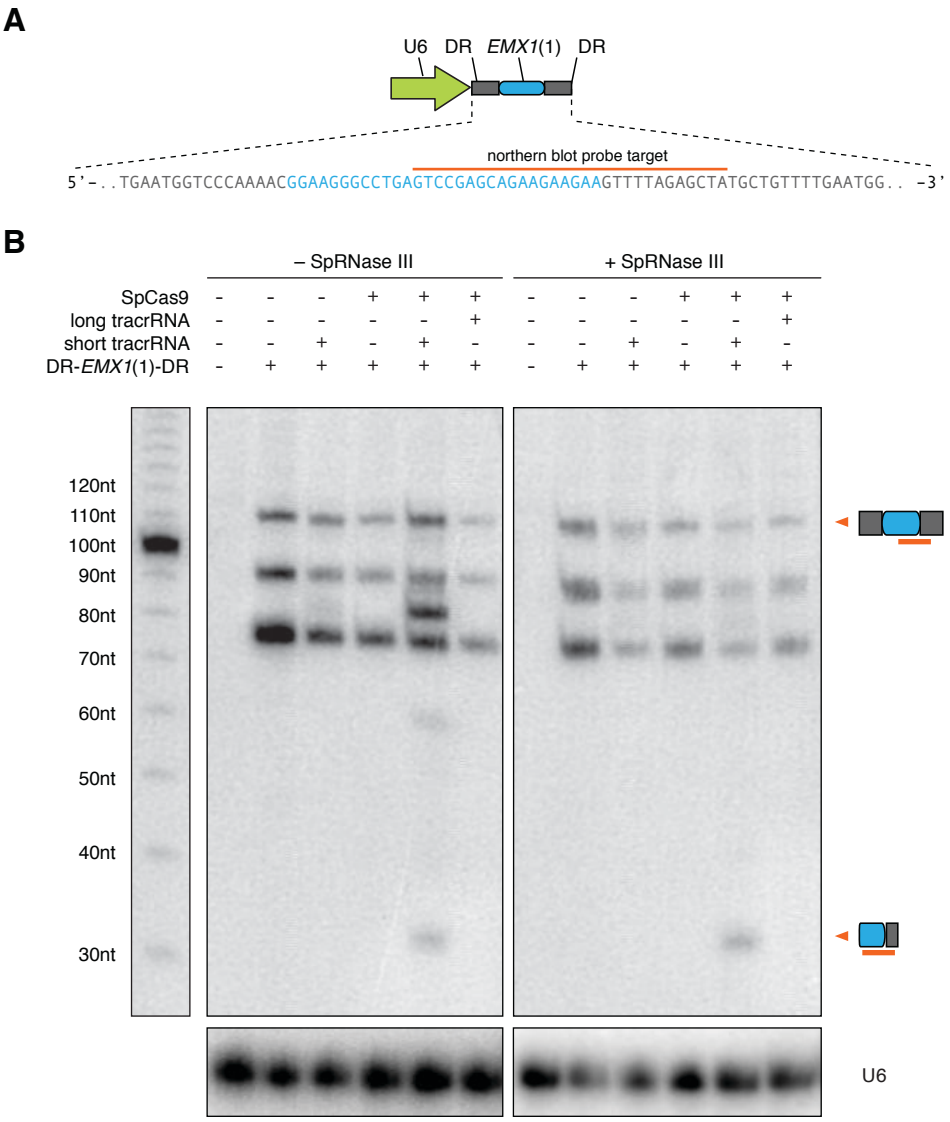


Fig. S4. Northern blot analysis of crRNA processing in mammalian cells. (A) Schematic showing the expression vector for a single spacer flanked by two direct repeats (DR-*EMX1(1)-DR*). The 30bp spacer targeting the human *EMX1* locus protospacer 1 (Table S1) is shown in blue and direct repeats are in shown in gray. Orange line indicates the region whose reverse-complement sequence is used to generate northern blot probes for *EMX1*(1) crRNA detection. (**B**) Northern blot analysis of total RNA extracted from 293FT cells transfected with U6 expression constructs carrying DR-*EMX1*(1)-DR. Left and right panels are from 293FT cells transfected without or with SpRNase III

(Fig. S4, continued) respectively. DR-*EMX1*(1)-DR was processed into mature crRNAs only in the presence of SpCas9 and short tracrRNA, and was not dependent on the presence of SpRNase III. The mature crRNA detected from transfected 293FT total RNA is ~33bp and is shorter than the 39-42bp mature crRNA from *S. pyogenes* (*124*), suggesting that the processed mature crRNA in human 293FT cells is likely different from the bacterial mature crRNA in *S. pyogenes*.

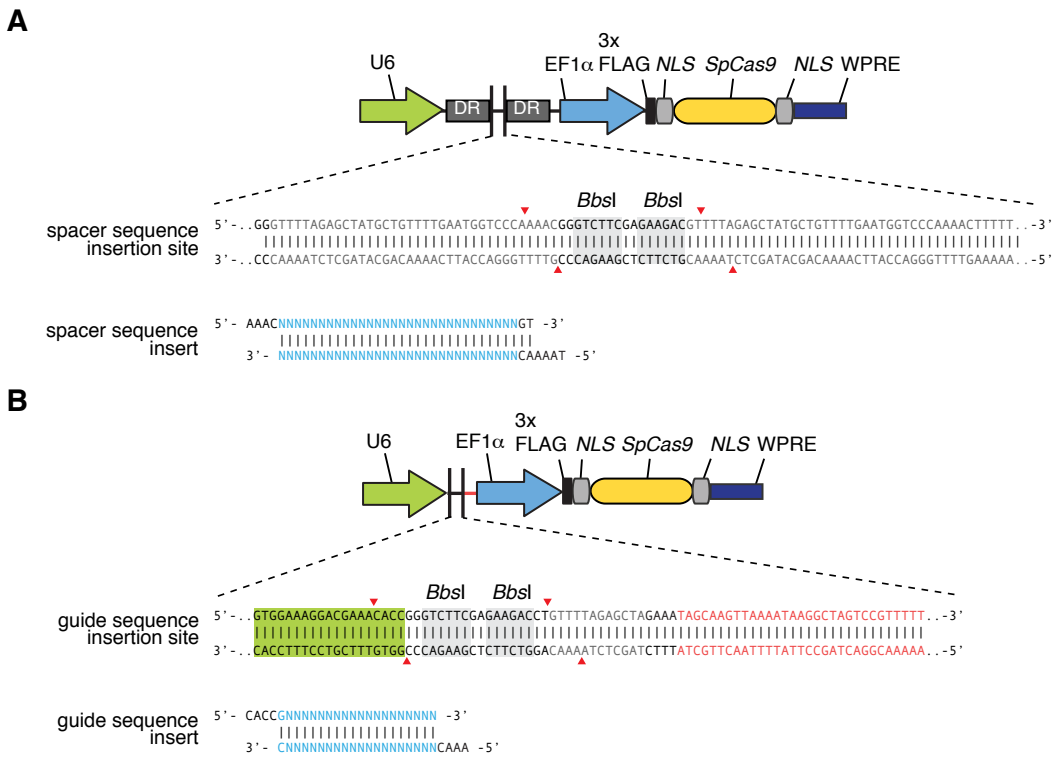


Fig. S5. Bicistronic expression vectors for pre-crRNA array or chimeric crRNA with Cas9. (A) Schematic showing the design of an expression vector for the pre-crRNA array. Spacers can be inserted between two *Bbs*I sites using annealed oligonucleotides. Sequence design for the oligonucleotides are shown below with the appropriate ligation adapters indicated. (B) Schematic of the expression vector for chimeric crRNA. The guide sequence can be inserted between two *Bbs*I sites using annealed oligonucleotides. The vector already contains the partial direct repeat (gray) and partial tracrRNA (red) sequences. WPRE, Woodchuck hepatitis virus post-transcriptional regulatory element.

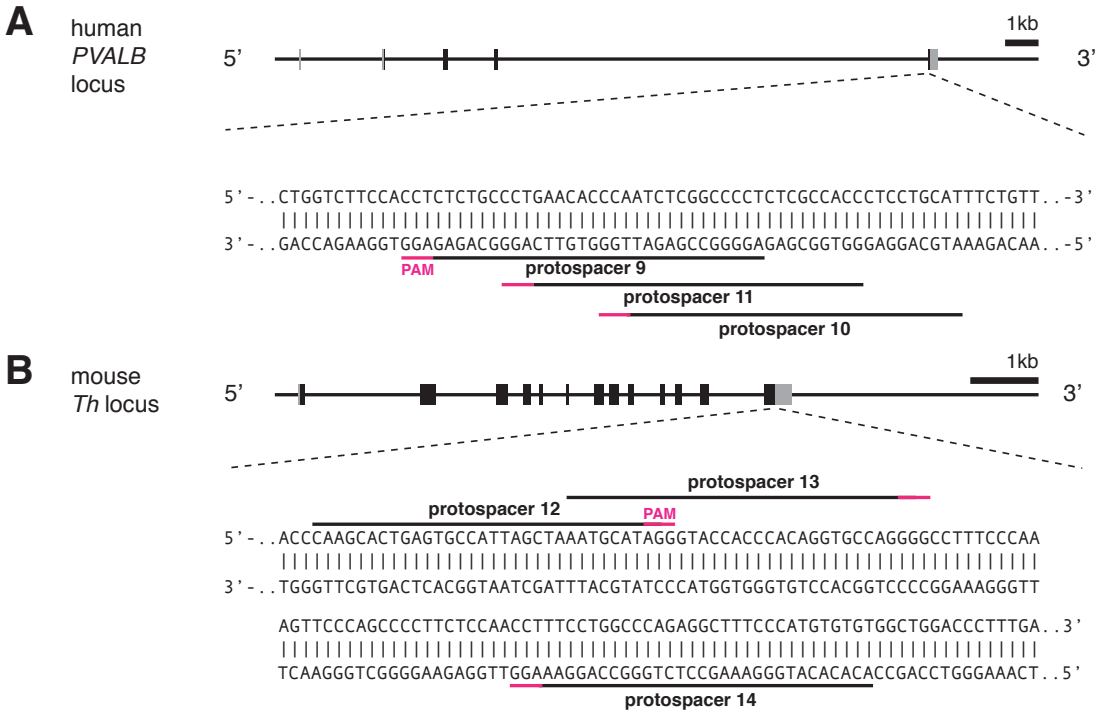
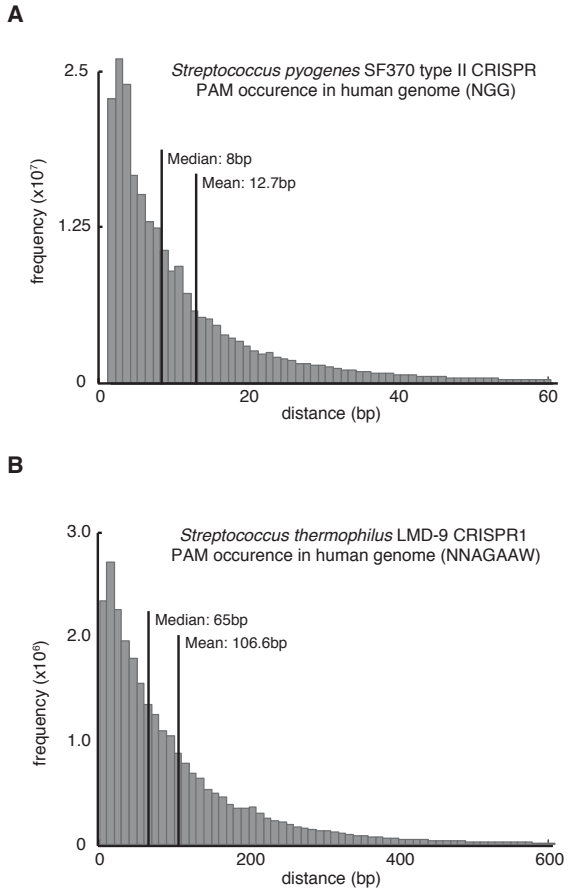


Fig. S6. Selection of protospacers in the human *PVALB* and mouse *Th* loci. Schematic of the human *PVALB* (A) and mouse *Th* (B) loci and the location of the three protospacers within the last exon of the *PVALB* and *Th* genes, respectively. The 30bp protospacers are indicated by black lines and the adjacent PAM sequences are indicated by the magenta bar. Protospacers on the sense and anti-sense strands are indicated above and below the DNA sequences respectively.



C

	NGG		NNAGAAW	
Chr	median	mean	median	mean
1	7	12.8	67	115.8
2	8	12.7	64	100.8
3	8	13.0	63	98.5
4	9	14.0	61	94.5
5	8	13.1	63	97.9
6	8	13.1	63	98.5
7	8	12.4	64	102.9
8	8	12.8	64	100.9
9	7	13.9	65	120.5
10	7	12.1	66	107.0
11	7	12.0	65	105.8
12	8	12.4	65	103.5
13	8	13.6	62	94.6
14	8	12.0	65	101.5
15	7	11.5	68	107.7
16	7	11.7	74	136.8
17	6	10.3	76	127.9
18	8	13.4	63	101.8
19	6	9.4	82	145.4
20	7	11.1	72	121.8
21	7	13.4	64	111.4
22	6	9.2	85	140.3
X	8	13.2	63	99.0
Y	8	29.2	62	223.7

Fig. S7. Occurrences of PAM sequences in the human genome. Histograms of distances between adjacent *Streptococcus pyogenes* SF370 type II CRISPR PAM (NGG) (A) and *Streptococcus thermophiles* LMD-9 CRISPR1 PAM (NNAGAAW) (B) in the human genome. (C) Distances for each PAM by chromosome. Chr, chromosome. Putative targets were identified using both the plus and minus strands of human chromosomal sequences. Given that there may be chromatin, DNA methylation-, RNA structure, and other factors that may limit the cleavage activity at some protospacer targets, it is important to note that the actual targeting ability might be less than the result of this computational analysis.

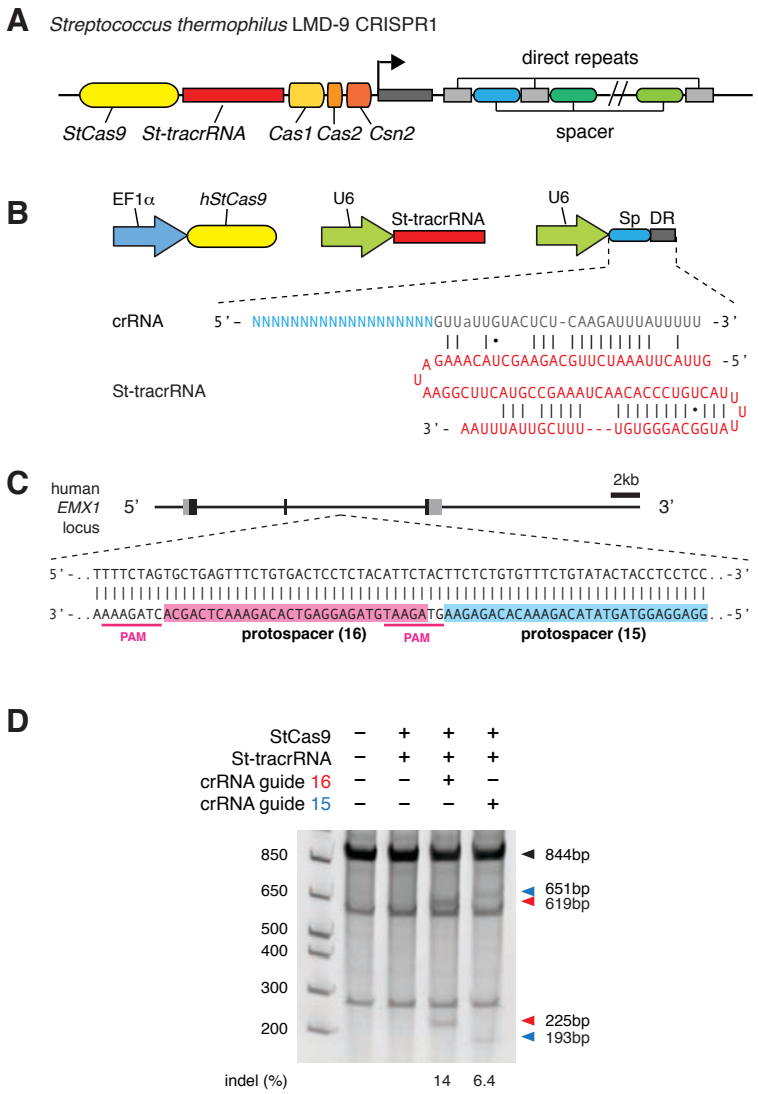


Fig S8. Type II CRISPR from *Streptococcus thermophilus* LMD-9 can also function in eukaryotic cells. (A) Schematic of CRISPR locus 2 from *Streptococcus thermophilus* LMD-9. (B) Design of the expression system for the *S. thermphilus* CRISPR system. Human codon-optimized *StCas9* is expressed using a constitutive EF1a promoter. Mature versions of tracrRNA and crRNA are expressed using the U6 promoter to ensure precise transcription initiation. Sequences for the mature crRNA and tracrRNA are shown. A single based indicated by the lower case "a" in the crRNA sequence was used to remove the polyU sequence, which serves as a RNA Pol III transcriptional terminator. Sp, spacer.

(Fig. S8, continued) (**C**) Schematic showing protospacer and corresponding PAM sequences targets in the human *EMX1* locus. Two protospacer sequences are highlighted and their corresponding PAM sequences satisfying the NNAGAAW motif are indicated by magenta lines. Both protospacers are targeting the anti-sense strand. (**D**) SURVEYOR assay showing StCas9-mediated cleavage in the target locus. RNA guide spacers 1 and 2 induced 14% and 6.4% respectively. Statistical analysis of cleavage activity across biological replica at these two protospacer sites can be found in Table S1.

Table S1. Protospacer sequences and modification efficiencies of mammalian genomic targets. Protospacer targets designed based on *Streptococcus pyogenes* type II CRISPR and *Streptococcus thermophilus* CRISPR1 loci with their requisite PAMs against three different genes in human and mouse genomes. Cells were transfected with Cas9 and either pre-crRNA/tracrRNA or chimeric RNA. Cells were analyzed 72 hours after transfection. Percent indels are calculated based on SURVEYOR assay results from indicated cell lines, $N = 3$ for all protospacer targets, errors are S.E.M. N.D., not detectable using the SURVEYOR assay; N.T., not tested in this study.

Cas9	target species	gene	protospacer ID	protospacer sequence (5' to 3')	PAM	strand	cell line tested	% indel (pre-crRNA + tracrRNA)	% indel (chimeric RNA)
S. pyogenes SF370 type II CRISPR	*Homo sapiens*	*EMX1*	1	GGAAGGGCCTGAGTCCGAGCAGAAGAAGAA	GGG	+	293FT	20 ± 1.8	6.7 ± 0.62
		EMX1	2	CATTGGAGGTGACATCGATGTCCTCCCCAT	TGG	–	293FT	2.1 ± 0.31	N.D.
		EMX1	3	GGACATCGATGTCACCTCCAATGACTAGGG	TGG	+	293FT	14 ± 1.1	N.D.
		EMX1	4	CATCGATGTCCTCCCCATTGGCCTGCTTCG	TGG	–	293FT	11 ± 1.7	N.D.
		EMX1	5	TTCGTGGCAATGCGCCACCGGTTGATGTGA	TGG	–	293FT	4.3 ± 0.46	2.1 ± 0.51
		EMX1	6	TCGTGGCAATGCGCCACCGGTTGATGTGAT	GGG	–	293FT	4.0 ± 0.66	0.41 ± 0.25
		EMX1	7	TCCAGCTTCTGCCGTTTGTACTTTGTCCTC	CGG	–	293FT	1.5 ± 0.12	N.D.
		EMX1	8	GGAGGGAGGGGCACAGATGAGAAACTCAGG	AGG	–	293FT	7.8 ± 0.83	2.3 ± 1.2
	Homo sapiens	*PVALB*	9	AGGGGCCGAGATTGGGTGTTCAGGGCAGAG	AGG	+	293FT	21 ± 2.6	6.5 ± 0.32
		PVALB	10	ATGCAGGAGGGTGGCGAGAGGGGCCGAGAT	TGG	+	293FT	N.D.	N.D.
		PVALB	11	GGTGGCGAGAGGGGCCGAGATTGGGTGTTC	AGG	+	293FT	N.D.	N.D.
	Mus musculus	*Th*	12	CAAGCACTGAGTGCCATTAGCTAAATGCAT	AGG	–	Neuro2A	27 ± 4.3	4.1 ± 2.2
		Th	13	AATGCATAGGGTACCACCCACAGGTGCCAG	GGG	–	Neuro2A	4.8 ± 1.2	N.D.
		Th	14	ACACACATGGGAAAGCCTCTGGGCCAGGAA	AGG	+	Neuro2A	11.3 ± 1.3	N.D.
S. thermophilus LMD-9 CRISPR1	*Homo sapiens*	*EMX1*	15	GGAGGAGGTAGTATACAGAAACACAGAGAA	GTAGAAT	–	293FT	14 ± 0.88	N.T.
		EMX1	16	AGAATGTAGAGGAGTCACAGAAACTCAGCA	CTAGAAA	–	293FT	7.8 ± 0.77	N.T.

Table S2. Sequences for primers and probes used for SURVEYOR assay, RFLP assay, genomic sequencing, and Northern blot.

Primer name	**Assay**	**Genomic Target**	**Primer sequence**
Sp-EMX1-F	SURVEYOR. sequencing	*EMX1*	AAAACCACCCTTCTCTCTGGC
Sp-EMX1-R	SURVEYOR. sequencing	*EMX1*	GGAGATTGGAGACACGGAGAG
Sp-PVALB-F	SURVEYOR. sequencing	*PVALB*	CTGGAAAGCCAATGCCTGAC
Sp-PVALB-R	SURVEYOR. sequencing	*PVALB*	GGCAGCAAACTCCTTGTCCT
Sp-Th-F	SURVEYOR. sequencing	*Th*	GTGCTTTGCAGAGGCCTACC
Sp-Th-R	SURVEYOR. sequencing	*Th*	CCTGGAGCGCATGCAGTAGT
St-EMX1-F	SURVEYOR. sequencing	*EMX1*	ACCTTCTGTGTTTCCACCATTC
St-EMX1-R	SURVEYOR. sequencing	*EMX1*	TTGGGGAGTGCACAGACTTC
Sp-EMX1-RFLP-F	RFLP, sequencing	*EMX1*	GGCTCCCTGGGTTCAAAGTA
Sp-EMX1-RFLP-R	RFLP, sequencing	*EMX1*	AGAGGGGTCTGGATGTCGTAA
Pb_EMX1_sp1	Northern Blot Probe	Not applicable	TAGCTCTAAAACTTCTTCTTCTGCTCGGAC
Pb_tracrRNA	Northern Blot Probe	Not applicable	CTAGCCTTATTTTAACTTGCTATGCTGTTT

Supplementary sequences

> U6-short tracrRNA (*Streptococcus pyogenes* SF370)

GAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGA
GATAATTGGAATTAATTTGACTGTAAACACAAAGATATTAGTACAAAATACGTGACGTA
GAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAATGGACTAT
CATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAA
AGGACGAAACACCGGAACCATTCAAAACAGCATAGCAAGTTAAAATAAGGCTAGTCCGT
TATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT

> U6-long tracrRNA (*Streptococcus pyogenes* SF370)

GAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGA
GATAATTGGAATTAATTTGACTGTAAACACAAAGATATTAGTACAAAATACGTGACGTA
GAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAATGGACTAT
CATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAA
AGGACGAAACACCGGTAGTATTAAGTATTGTTTTATGGCTGATAAATTTCTTTGAATTT
CTCCTTGATTATTTGTTATAAAAGTTATAAAATAATCTTGTTGGAACCATTCAAAACAG
CATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG
TGCTTTTTTT

> U6-DR-BbsI backbone-DR (*Streptococcus pyogenes* SF370)

GAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGA
GATAATTGGAATTAATTTGACTGTAAACACAAAGATATTAGTACAAAATACGTGACGTA
GAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAATGGACTAT
CATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAA
AGGACGAAACACCGGGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACGGGTCTTC
GAGAAGACGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC

> U6-chimeric RNA-BbsI backbone (*Streptococcus pyogenes* SF370)

GAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGA
GATAATTGGAATTAATTTGACTGTAAACACAAAGATATTAGTACAAAATACGTGACGTA
GAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAATGGACTAT
CATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAA
AGGACGAAACACCGGGTCTTCGAGAAGACCTGTTTTAGAGCTAGAAATAGCAAGTTAAA
ATAAGGCTAGTCCG

> 3xFLAG-NLS-SpCas9-NLS

ATGGACTATAAGGACCACGACGGAGACTACAAGGATCATGATATTGATTACAAAGACGA
TGACGATAAGATGGCCCCAAAGAAGAAGCGGAAGGTCGGTATCCACGGAGTCCCAGCAG
CCGACAAGAAGTACAGCATCGGCCTGGACATCGGCACCAACTCTGTGGGCTGGGCCGTG
ATCACCGACGAGTACAAGGTGCCCAGCAAGAAATTCAAGGTGCTGGGCAACACCGACCG
GCACAGCATCAAGAAGAACCTGATCGGAGCCCTGCTGTTCGACAGCGGCGAAACAGCCG
AGGCCACCCGGCTGAAGAGAACCGCCAGAAGAAGATACACCAGACGGAAGAACCGGATC
TGCTATCTGCAAGAGATCTTCAGCAACGAGATGGCCAAGGTGGACGACAGCTTCTTCCA
CAGACTGGAAGAGTCCTTCCTGGTGGAAGAGGATAAGAAGCACGAGCGGCACCCCATCT
TCGGCAACATCGTGGACGAGGTGGCCTACCACGAGAAGTACCCCACCATCTACCACCTG
AGAAAGAAACTGGTGGACAGCACCGACAAGGCCGACCTGCGGCTGATCTATCTGGCCCT
GGCCCACATGATCAAGTTCCGGGGCCACTTCCTGATCGAGGGCGACCTGAACCCCGACA
ACAGCGACGTGGACAAGCTGTTCATCCAGCTGGTGCAGACCTACAACCAGCTGTTCGAG
GAAAACCCCATCAACGCCAGCGGCGTGGACGCCAAGGCCATCCTGTCTGCCAGACTGAG
CAAGAGCAGACGGCTGGAAAATCTGATCGCCCAGCTGCCCGGCGAGAAGAAGAATGGCC
TGTTCGGCAACCTGATTGCCCTGAGCCTGGGCCTGACCCCCAACTTCAAGAGCAACTTC
GACCTGGCCGAGGATGCCAAACTGCAGCTGAGCAAGGACACCTACGACGACGACCTGGA
CAACCTGCTGGCCCAGATCGGCGACCAGTACGCCGACCTGTTTCTGGCCGCCAAGAACC
TGTCCGACGCCATCCTGCTGAGCGACATCCTGAGAGTGAACACCGAGATCACCAAGGCC

CCCCTGAGCGCCTCTATGATCAAGAGATACGACGAGCACCACCAGGACCTGACCCTGCT
GAAAGCTCTCGTGCGGCAGCAGCTGCCTGAGAAGTACAAAGAGATTTTCTTCGACCAGA
GCAAGAACGGCTACGCCGGCTACATTGACGGCGGAGCCAGCCAGGAAGAGTTCTACAAG
TTCATCAAGCCCATCCTGGAAAAGATGGACGGCACCGAGGAACTGCTCGTGAAGCTGAA
CAGAGAGGACCTGCTGCGGAAGCAGCGGACCTTCGACAACGGCAGCATCCCCCACCAGA
TCCACCTGGGAGAGCTGCACGCCATTCTGCGGCGGCAGGAAGATTTTTACCCATTCCTG
AAGGACAACCGGGAAAAGATCGAGAAGATCCTGACCTTCCGCATCCCCTACTACGTGGG
CCCTCTGGCCAGGGGAAACAGCAGATTCGCCTGGATGACCAGAAAGAGCGAGGAAACCA
TCACCCCCTGGAACTTCGAGGAAGTGGTGGACAAGGGCGCTTCCGCCCAGAGCTTCATC
GAGCGGATGACCAACTTCGATAAGAACCTGCCCAACGAGAAGGTGCTGCCCAAGCACAG
CCTGCTGTACGAGTACTTCACCGTGTATAACGAGCTGACCAAAGTGAAATACGTGACCG
AGGGAATGAGAAAGCCCGCCTTCCTGAGCGGCGAGCAGAAAAAGGCCATCGTGGACCTG
CTGTTCAAGACCAACCGGAAAGTGACCGTGAAGCAGCTGAAAGAGGACTACTTCAAGAA
AATCGAGTGCTTCGACTCCGTGGAAATCTCCGGCGTGGAAGATCGGTTCAACGCCTCCC
TGGGCACATACCACGATCTGCTGAAAATTATCAAGGACAAGGACTTCCTGGACAATGAG
GAAAACGAGGACATTCTGGAAGATATCGTGCTGACCCTGACACTGTTTGAGGACAGAGA
GATGATCGAGGAACGGCTGAAAACCTATGCCCACCTGTTCGACGACAAAGTGATGAAGC
AGCTGAAGCGGCGGAGATACACCGGCTGGGGCAGGCTGAGCCGGAAGCTGATCAACGGC
ATCCGGGACAAGCAGTCCGGCAAGACAATCCTGGATTTCCTGAAGTCCGACGGCTTCGC
CAACAGAAACTTCATGCAGCTGATCCACGACGACAGCCTGACCTTTAAAGAGGACATCC
AGAAAGCCCAGGTGTCCGGCCAGGGCGATAGCCTGCACGAGCACATTGCCAATCTGGCC
GGCAGCCCCGCCATTAAGAAGGGCATCCTGCAGACAGTGAAGGTGGTGGACGAGCTCGT
GAAAGTGATGGGCCGGCACAAGCCCGAGAACATCGTGATCGAAATGGCCAGAGAGAACC
AGACCACCCAGAAGGGACAGAAGAACAGCCGCGAGAGAATGAAGCGGATCGAAGAGGGC
ATCAAAGAGCTGGGCAGCCAGATCCTGAAAGAACACCCCGTGGAAAACACCCAGCTGCA
GAACGAGAAGCTGTACCTGTACTACCTGCAGAATGGGCGGGATATGTACGTGGACCAGG
AACTGGACATCAACCGGCTGTCCGACTACGATGTGGACCATATCGTGCCTCAGAGCTTT
CTGAAGGACGACTCCATCGACAACAAGGTGCTGACCAGAAGCGACAAGAACCGGGGCAA

GAGCGACAACGTGCCCTCCGAAGAGGTCGTGAAGAAGATGAAGAACTACTGGCGGCAGC
TGCTGAACGCCAAGCTGATTACCCAGAGAAAGTTCGACAATCTGACCAAGGCCGAGAGA
GGCGGCCTGAGCGAACTGGATAAGGCCGGCTTCATCAAGAGACAGCTGGTGGAAACCCG
GCAGATCACAAAGCACGTGGCACAGATCCTGGACTCCCGGATGAACACTAAGTACGACG
AGAATGACAAGCTGATCCGGGAAGTGAAAGTGATCACCCTGAAGTCCAAGCTGGTGTCC
GATTTCCGGAAGGATTTCCAGTTTTACAAAGTGCGCGAGATCAACAACTACCACCACGC
CCACGACGCCTACCTGAACGCCGTCGTGGGAACCGCCCTGATCAAAAAGTACCCTAAGC
TGGAAAGCGAGTTCGTGTACGGCGACTACAAGGTGTACGACGTGCGGAAGATGATCGCC
AAGAGCGAGCAGGAAATCGGCAAGGCTACCGCCAAGTACTTCTTCTACAGCAACATCAT
GAACTTTTTCAAGACCGAGATTACCCTGGCCAACGGCGAGATCCGGAAGCGGCCTCTGA
TCGAGACAAACGGCGAAACCGGGGAGATCGTGTGGGATAAGGGCCGGGATTTTGCCACC
GTGCGGAAAGTGCTGAGCATGCCCCAAGTGAATATCGTGAAAAAGACCGAGGTGCAGAC
AGGCGGCTTCAGCAAAGAGTCTATCCTGCCCAAGAGGAACAGCGATAAGCTGATCGCCA
GAAAGAAGGACTGGGACCCTAAGAAGTACGGCGGCTTCGACAGCCCCACCGTGGCCTAT
TCTGTGCTGGTGGTGGCCAAAGTGGAAAAGGGCAAGTCCAAGAAACTGAAGAGTGTGAA
AGAGCTGCTGGGGATCACCATCATGGAAAGAAGCAGCTTCGAGAAGAATCCCATCGACT
TTCTGGAAGCCAAGGGCTACAAAGAAGTGAAAAAGGACCTGATCATCAAGCTGCCTAAG
TACTCCCTGTTCGAGCTGGAAAACGGCCGGAAGAGAATGCTGGCCTCTGCCGGCGAACT
GCAGAAGGGAAACGAACTGGCCCTGCCCTCCAAATATGTGAACTTCCTGTACCTGGCCA
GCCACTATGAGAAGCTGAAGGGCTCCCCCGAGGATAATGAGCAGAAACAGCTGTTTGTG
GAACAGCACAAGCACTACCTGGACGAGATCATCGAGCAGATCAGCGAGTTCTCCAAGAG
AGTGATCCTGGCCGACGCTAATCTGGACAAAGTGCTGTCCGCCTACAACAAGCACCGGG
ATAAGCCCATCAGAGAGCAGGCCGAGAATATCATCCACCTGTTTACCCTGACCAATCTG
GGAGCCCCTGCCGCCTTCAAGTACTTTGACACCACCATCGACCGGAAGAGGTACACCAG
CACCAAAGAGGTGCTGGACGCCACCCTGATCCACCAGAGCATCACCGGCCTGTACGAGA
CACGGATCGACCTGTCTCAGCTGGGAGGCGACAAGCGTCCTGCTGCTACTAAGAAAGCT
GGTCAAGCTAAGAAAAAGAAA

> SpRNase3-mCherry-NLS

ATGAAGCAGCTGGAGGAGTTACTTTCTACCTCTTTCGACATCCAGTTTAATGACCTGAC
CCTGCTGGAAACCGCCTTCACTCACACCTCCTACGCGAATGAGCACCGCCTACTGAATG
TGAGCCACAACGAGCGCCTGGAGTTTCTGGGGGATGCTGTCTTACAGCTGATCATCTCT
GAATATCTGTTTGCCAAATACCCTAAGAAAACCGAAGGGGACATGTCAAAGCTGCGCTC
CATGATAGTCAGGGAAGAGAGCCTGGCGGGCTTTAGTCGTTTTTGCTCATTCGACGCTT
ATATCAAGCTGGGAAAAGGCGAAGAGAAGTCCGGCGGCAGGAGGCGCGATACAATTCTG
GGCGATCTCTTTGAAGCGTTTCTGGGCGCACTTCTACTGGACAAAGGGATCGACGCAGT
CCGCCGCTTTCTGAAACAAGTGATGATCCCTCAGGTCGAAAAGGGAAACTTCGAGAGAG
TGAAGGACTATAAAACATGTTTGCAGGAATTTCTCCAGACCAAGGGAGATGTAGCAATA
GATTATCAGGTAATAAGTGAGAAAGGACCAGCTCACGCCAAACAATTCGAAGTTAGCAT
CGTTGTTAATGGCGCAGTGTTGTCGAAGGGCTTGGGTAAATCAAAAAAACTGGCCGAGC
AGGACGCTGCTAAAAACGCCCTCGCTCAGCTCAGCGAGGTAGGATCCGTGAGCAAGGGC
GAGGAGGATAACATGGCCATCATCAAGGAGTTCATGCGCTTCAAGGTGCACATGGAGGG
CTCCGTGAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGAGGGCCGCCCCTACGAGG
GCACCCAGACCGCCAAGCTGAAGGTGACCAAGGGTGGCCCCCTGCCCTTCGCCTGGGAC
ATCCTGTCCCCTCAGTTCATGTACGGCTCCAAGGCCTACGTGAAGCACCCCGCCGACAT
CCCCGACTACTTGAAGCTGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGATGAACT
TCGAGGACGGCGGCGTGGTGACCGTGACCCAGGACTCCTCCCTGCAGGACGGCGAGTTC
ATCTACAAGGTGAAGCTGCGCGGCACCAACTTCCCCTCCGACGGCCCCGTAATGCAGAA
GAAGACCATGGGCTGGGAGGCCTCCTCCGAGCGGATGTACCCCGAGGACGGCGCCCTGA
AGGGCGAGATCAAGCAGAGGCTGAAGCTGAAGGACGGCGGCCACTACGACGCTGAGGTC
AAGACCACCTACAAGGCCAAGAAGCCCGTGCAGCTGCCCGGCGCCTACAACGTCAACAT
CAAGTTGGACATCACCTCCCACAACGAGGACTACACCATCGTGGAACAGTACGAACGCG
CCGAGGGCCGCCACTCCACCGGCGGCATGGACGAGCTGTACAAGAAGCGTCCTGCTGCT
ACTAAGAAAGCTGGTCAAGCTAAGAAAAAGAAA

> 3xFLAG-NLS-SpCas9n-NLS (the D10A nickase mutation is labeled in red)

ATGGACTATAAGGACCACGACGGAGACTACAAGGATCATGATATTGATTACAAAGACGA
TGACGATAAGATGGCCCCAAAGAAGAAGCGGAAGGTCGGTATCCACGGAGTCCCAGCAG
CCGACAAGAAGTACAGCATCGGCCTGGCCATCGGCACCAACTCTGTGGGCTGGGCCGTG
ATCACCGACGAGTACAAGGTGCCCAGCAAGAAATTCAAGGTGCTGGGCAACACCGACCG
GCACAGCATCAAGAAGAACCTGATCGGAGCCCTGCTGTTCGACAGCGGCGAAACAGCCG
AGGCCACCCGGCTGAAGAGAACCGCCAGAAGAAGATACACCAGACGGAAGAACCGGATC
TGCTATCTGCAAGAGATCTTCAGCAACGAGATGGCCAAGGTGGACGACAGCTTCTTCCA
CAGACTGGAAGAGTCCTTCCTGGTGGAAGAGGATAAGAAGCACGAGCGGCACCCCATCT
TCGGCAACATCGTGGACGAGGTGGCCTACCACGAGAAGTACCCCACCATCTACCACCTG
AGAAAGAAACTGGTGGACAGCACCGACAAGGCCGACCTGCGGCTGATCTATCTGGCCCT
GGCCCACATGATCAAGTTCCGGGGCCACTTCCTGATCGAGGGCGACCTGAACCCCGACA
ACAGCGACGTGGACAAGCTGTTCATCCAGCTGGTGCAGACCTACAACCAGCTGTTCGAG
GAAAACCCCATCAACGCCAGCGGCGTGGACGCCAAGGCCATCCTGTCTGCCAGACTGAG
CAAGAGCAGACGGCTGGAAAATCTGATCGCCCAGCTGCCCGGCGAGAAGAAGAATGGCC
TGTTCGGCAACCTGATTGCCCTGAGCCTGGGCCTGACCCCCAACTTCAAGAGCAACTTC
GACCTGGCCGAGGATGCCAAACTGCAGCTGAGCAAGGACACCTACGACGACGACCTGGA
CAACCTGCTGGCCCAGATCGGCGACCAGTACGCCGACCTGTTTCTGGCCGCCAAGAACC
TGTCCGACGCCATCCTGCTGAGCGACATCCTGAGAGTGAACACCGAGATCACCAAGGCC
CCCCTGAGCGCCTCTATGATCAAGAGATACGACGAGCACCACCAGGACCTGACCCTGCT
GAAAGCTCTCGTGCGGCAGCAGCTGCCTGAGAAGTACAAAGAGATTTTCTTCGACCAGA
GCAAGAACGGCTACGCCGGCTACATTGACGGCGGAGCCAGCCAGGAAGAGTTCTACAAG
TTCATCAAGCCCATCCTGGAAAAGATGGACGGCACCGAGGAACTGCTCGTGAAGCTGAA
CAGAGAGGACCTGCTGCGGAAGCAGCGGACCTTCGACAACGGCAGCATCCCCCACCAGA
TCCACCTGGGAGAGCTGCACGCCATTCTGCGGCGGCAGGAAGATTTTTACCCATTCCTG
AAGGACAACCGGGAAAAGATCGAGAAGATCCTGACCTTCCGCATCCCCTACTACGTGGG
CCCTCTGGCCAGGGGAAACAGCAGATTCGCCTGGATGACCAGAAAGAGCGAGGAAACCA

TCACCCCCTGGAACTTCGAGGAAGTGGTGGACAAGGGCGCTTCCGCCCAGAGCTTCATC
GAGCGGATGACCAACTTCGATAAGAACCTGCCCAACGAGAAGGTGCTGCCCAAGCACAG
CCTGCTGTACGAGTACTTCACCGTGTATAACGAGCTGACCAAAGTGAAATACGTGACCG
AGGGAATGAGAAAGCCCGCCTTCCTGAGCGGCGAGCAGAAAAAGGCCATCGTGGACCTG
CTGTTCAAGACCAACCGGAAAGTGACCGTGAAGCAGCTGAAAGAGGACTACTTCAAGAA
AATCGAGTGCTTCGACTCCGTGGAAATCTCCGGCGTGGAAGATCGGTTCAACGCCTCCC
TGGGCACATACCACGATCTGCTGAAAATTATCAAGGACAAGGACTTCCTGGACAATGAG
GAAAACGAGGACATTCTGGAAGATATCGTGCTGACCCTGACACTGTTTGAGGACAGAGA
GATGATCGAGGAACGGCTGAAAACCTATGCCCACCTGTTCGACGACAAAGTGATGAAGC
AGCTGAAGCGGCGGAGATACACCGGCTGGGGCAGGCTGAGCCGGAAGCTGATCAACGGC
ATCCGGGACAAGCAGTCCGGCAAGACAATCCTGGATTTCCTGAAGTCCGACGGCTTCGC
CAACAGAAACTTCATGCAGCTGATCCACGACGACAGCCTGACCTTTAAAGAGGACATCC
AGAAAGCCCAGGTGTCCGGCCAGGGCGATAGCCTGCACGAGCACATTGCCAATCTGGCC
GGCAGCCCCGCCATTAAGAAGGGCATCCTGCAGACAGTGAAGGTGGTGGACGAGCTCGT
GAAAGTGATGGGCCGGCACAAGCCCGAGAACATCGTGATCGAAATGGCCAGAGAGAACC
AGACCACCCAGAAGGGACAGAAGAACAGCCGCGAGAGAATGAAGCGGATCGAAGAGGGC
ATCAAAGAGCTGGGCAGCCAGATCCTGAAAGAACACCCCGTGGAAAACACCCAGCTGCA
GAACGAGAAGCTGTACCTGTACTACCTGCAGAATGGGCGGGATATGTACGTGGACCAGG
AACTGGACATCAACCGGCTGTCCGACTACGATGTGGACCATATCGTGCCTCAGAGCTTT
CTGAAGGACGACTCCATCGACAACAAGGTGCTGACCAGAAGCGACAAGAACCGGGGCAA
GAGCGACAACGTGCCCTCCGAAGAGGTCGTGAAGAAGATGAAGAACTACTGGCGGCAGC
TGCTGAACGCCAAGCTGATTACCCAGAGAAAGTTCGACAATCTGACCAAGGCCGAGAGA
GGCGGCCTGAGCGAACTGGATAAGGCCGGCTTCATCAAGAGACAGCTGGTGGAAACCCG
GCAGATCACAAAGCACGTGGCACAGATCCTGGACTCCCGGATGAACACTAAGTACGACG
AGAATGACAAGCTGATCCGGGAAGTGAAAGTGATCACCCTGAAGTCCAAGCTGGTGTCC
GATTTCCGGAAGGATTTCCAGTTTTACAAAGTGCGCGAGATCAACAACTACCACCACGC
CCACGACGCCTACCTGAACGCCGTCGTGGGAACCGCCCTGATCAAAAAGTACCCTAAGC
TGGAAAGCGAGTTCGTGTACGGCGACTACAAGGTGTACGACGTGCGGAAGATGATCGCC

AAGAGCGAGCAGGAAATCGGCAAGGCTACCGCCAAGTACTTCTTCTACAGCAACATCAT
GAACTTTTTCAAGACCGAGATTACCCTGGCCAACGGCGAGATCCGGAAGCGGCCTCTGA
TCGAGACAAACGGCGAAACCGGGGAGATCGTGTGGGATAAGGGCCGGGATTTTGCCACC
GTGCGGAAAGTGCTGAGCATGCCCCAAGTGAATATCGTGAAAAAGACCGAGGTGCAGAC
AGGCGGCTTCAGCAAAGAGTCTATCCTGCCCAAGAGGAACAGCGATAAGCTGATCGCCA
GAAAGAAGGACTGGGACCCTAAGAAGTACGGCGGCTTCGACAGCCCCACCGTGGCCTAT
TCTGTGCTGGTGGTGGCCAAAGTGGAAAAGGGCAAGTCCAAGAAACTGAAGAGTGTGAA
AGAGCTGCTGGGGATCACCATCATGGAAAGAAGCAGCTTCGAGAAGAATCCCATCGACT
TTCTGGAAGCCAAGGGCTACAAAGAAGTGAAAAAGGACCTGATCATCAAGCTGCCTAAG
TACTCCCTGTTCGAGCTGGAAAACGGCCGGAAGAGAATGCTGGCCTCTGCCGGCGAACT
GCAGAAGGGAAACGAACTGGCCCTGCCCTCCAAATATGTGAACTTCCTGTACCTGGCCA
GCCACTATGAGAAGCTGAAGGGCTCCCCCGAGGATAATGAGCAGAAACAGCTGTTTGTG
GAACAGCACAAGCACTACCTGGACGAGATCATCGAGCAGATCAGCGAGTTCTCCAAGAG
AGTGATCCTGGCCGACGCTAATCTGGACAAAGTGCTGTCCGCCTACAACAAGCACCGGG
ATAAGCCCATCAGAGAGCAGGCCGAGAATATCATCCACCTGTTTACCCTGACCAATCTG
GGAGCCCCTGCCGCCTTCAAGTACTTTGACACCACCATCGACCGGAAGAGGTACACCAG
CACCAAAGAGGTGCTGGACGCCACCCTGATCCACCAGAGCATCACCGGCCTGTACGAGA
CACGGATCGACCTGTCTCAGCTGGGAGGCGACAAGCGTCCTGCTGCTACTAAGAAAGCT
GGTCAAGCTAAGAAAAAGAAA

> hEMX1-HRTemplate-*Hin*dIII-*Nhe*I

GAATGCTGCCCTCAGACCCGCTTCCTCCCTGTCCTTGTCTGTCCAAGGAGAATGAGGTC
TCACTGGTGGATTTCGGACTACCCTGAGGAGCTGGCACCTGAGGGACAAGGCCCCCCAC
CTGCCCAGCTCCAGCCTCTGATGAGGGGTGGGAGAGAGCTACATGAGGTTGCTAAGAAA
GCCTCCCCTGAAGGAGACCACACAGTGTGTGAGGTTGGAGTCTCTAGCAGCGGGTTCTG
TGCCCCCAGGGATAGTCTGGCTGTCCAGGCACTGCTCTTGATATAAACACCACCTCCTA
GTTATGAAACCATGCCCATTCTGCCTCTCTGTATGGAAAAGAGCATGGGGCTGGCCCGT
GGGGTGGTGTCCACTTTAGGCCCTGTGGGAGATCATGGGAACCCACGCAGTGGGTCATA

GGCTCTCTCATTTACTACTCACATCCACTCTGTGAAGAAGCGATTATGATCTCTCCTCT
AGAAACTCGTAGAGTCCCATGTCTGCCGGCTTCCAGAGCCTGCACTCCTCCACCTTGGC
TTGGCTTTGCTGGGGCTAGAGGAGCTAGGATGCACAGCAGCTCTGTGACCCTTTGTTTG
AGAGGAACAGGAAAACCACCCTTCTCTCTGGCCCACTGTGTCCTCTTCCTGCCCTGCCA
TCCCCTTCTGTGAATGTTAGACCCATGGGAGCAGCTGGTCAGAGGGGACCCCGGCCTGG
GGCCCCTAACCCTATGTAGCCTCAGTCTTCCCATCAGGCTCTCAGCTCAGCCTGAGTGT
TGAGGCCCCAGTGGCTGCTCTGGGGGCCTCCTGAGTTTCTCATCTGTGCCCCTCCCTCC
CTGGCCCAGGTGAAGGTGTGGTTCCAGAACCGGAGGACAAAGTACAAACGGCAGAAGCT
GGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGAAGAAGGGCTCCCATCACATCAACCGGT
GGCGCATTGCCACGAAGCAGGCCAATGGGGAGGACATCGATGTCACCTCCAATGACaag
cttgctagcGGTGGGCAACCACAAACCCACGAGGGCAGAGTGCTGCTTGCTGCTGGCCA
GGCCCCTGCGTGGGCCCAAGCTGGACTCTGGCCACTCCCTGGCCAGGCTTTGGGGAGGC
CTGGAGTCATGGCCCCACAGGGCTTGAAGCCCGGGGCCGCCATTGACAGAGGGACAAGC
AATGGGCTGGCTGAGGCCTGGGACCACTTGGCCTTCTCCTCGGAGAGCCTGCCTGCCTG
GGCGGGCCCGCCCGCCACCGCAGCCTCCCAGCTGCTCTCCGTGTCTCCAATCTCCCTTT
TGTTTTGATGCATTTCTGTTTTAATTTATTTTCCAGGCACCACTGTAGTTTAGTGATCC
CCAGTGTCCCCCTTCCCTATGGGAATAATAAAAGTCTCTCTCTTAATGACACGGGCATC
CAGCTCCAGCCCCAGAGCCTGGGGTGGTAGATTCCGGCTCTGAGGGCCAGTGGGGGCTG
GTAGAGCAAACGCGTTCAGGGCCTGGGAGCCTGGGGTGGGGTACTGGTGGAGGGGGTCA
AGGGTAATTCATTAACTCCTCTCTTTTGTTGGGGGACCCTGGTCTCTACCTCCAGCTCC
ACAGCAGGAGAAACAGGCTAGACATAGGGAAGGGCCATCCTGTATCTTGAGGGAGGACA
GGCCCAGGTCTTTCTTAACGTATTGAGAGGTGGGAATCAGGCCCAGGTAGTTCAATGGG
AGAGGGAGAGTGCTTCCCTCTGCCTAGAGACTCTGGTGGCTTCTCCAGTTGAGGAGAAA
CCAGAGGAAAGGGGAGGATTGGGGTCTGGGGGAGGGAACACCATTCACAAAGGCTGACG
GTTCCAGTCCGAAGTCGTGGGCCCACCAGGATGCTCACCTGTCCTTGGAGAACCGCTGG
GCAGGTTGAGACTGCAGAGACAGGGCTTAAGGCTGAGCCTGCAACCAGTCCCCAGTGAC
TCAGGGCCTCCTCAGCCCAAGAAAGAGCAACGTGCCAGGGCCCGCTGAGCTCTTGTGTT
CACCTG

> NLS-StCsn1-NLS

ATGAAAAGGCCGGCGGCCACGAAAAAGGCCGGCCAGGCAAAAAAGAAAAAGTCCGACCT
GGTACTTGGACTGGATATTGGTATCGGTTCGGTGGGAGTCGGAATCCTCAACAAGGTCA
CGGGGGAGATCATTCACAAGAACTCGCGGATCTTCCCCGCAGCTCAGGCTGAGAACAAC
TTGGTGCGGAGAACGAATAGGCAGGGCAGGCGACTGGCGAGGAGGAAGAAACACAGGAG
AGTCCGATTGAACCGGCTGTTCGAGGAGTCCGGTTTGATCACCGACTTTACGAAAATCT
CGATTAACCTTAATCCCTATCAGCTTCGGGTGAAAGGCCTGACAGACGAACTTTCGAAT
GAGGAACTTTTCATCGCGCTGAAAAACATGGTCAAGCACAGAGGGATTTCCTACCTCGA
TGACGCCTCGGATGACGGAAATTCCTCAGTAGGAGATTATGCACAGATCGTGAAAGAGA
ACTCAAAGCAACTGGAAACAAAGACACCGGGGCAGATCCAACTTGAAAGATACCAGACA
TACGGACAGCTCAGAGGAGATTTTACGGTGGAGAAGGACGGTAAAAAGCACAGACTCAT
TAACGTATTTCCCACGTCGGCGTACAGATCCGAAGCGCTCCGCATCCTTCAGACTCAAC
AGGAGTTCAACCCGCAAATTACTGATGAGTTCATCAACCGCTATTTGGAAATCTTGACC
GGAAAGCGCAAGTATTATCATGGGCCGGGTAATGAGAAATCCAGAACAGATTACGGCCG
ATACAGAACTTCGGGGGAAACCTTGGATAACATCTTTGGTATTTTGATTGGAAAGTGCA
CCTTTTACCCGGACGAGTTTCGAGCGGCCAAGGCGTCATACACAGCACAAGAGTTTAAT
CTCTTGAATGATTTGAACAACTTGACGGTCCCCACGGAGACAAAGAAGCTCTCCAAAGA
GCAAAAGAACCAAATCATCAACTACGTCAAGAACGAGAAGGCTATGGGGCCAGCGAAGC
TGTTCAAGTATATCGCTAAACTTCTCAGCTGTGATGTGGCGGACATCAAAGGGTACCGA
ATCGACAAGTCGGGAAAAGCGGAAATTCACACGTTTGAAGCATATCGAAAGATGAAAAC
GTTGGAAACACTGGACATTGAGCAGATGGACCGGGAAACGCTCGACAAACTGGCATACG
TGCTCACGTTGAATACTGAACGAGAGGGAATCCAAGAGGCCCTTGAACATGAGTTCGCC
GATGGATCGTTCAGCCAGAAGCAGGTCGACGAACTTGTGCAATTCCGCAAGGCGAATAG
CTCCATCTTCGGGAAGGGATGGCACAACTTTTCGGTCAAACTCATGATGGAGTTGATCC
CAGAACTTTATGAGACTTCGGAGGAGCAAATGACGATCTTGACGCGCTTGGGGAAACAG
AAAACGACAAGCTCATCGAACAAAACTAAGTACATTGATGAGAAATTGCTGACGGAAGA
AATCTATAATCCGGTAGTAGCGAAATCGGTAAGACAAGCGATCAAAATCGTGAACGCGG

CGATCAAGGAATATGGTGACTTTGATAACATCGTAATTGAAATGGCTAGAGAGACGAAC

GAAGATGACGAGAAAAAGGCAATCCAGAAGATCCAGAAGGCCAACAAGGATGAAAAAGA

TGCAGCGATGCTTAAAGCGGCCAACCAATACAATGGAAAGGCGGAGCTGCCCCATTCAG

TGTTTCACGGTCATAAACAGTTGGCGACCAAGATCCGACTCTGGCATCAGCAGGGTGAG

CGGTGTCTCTACACCGGAAAGACTATCTCCATCCATGACTTGATTAACAATTCGAACCA

GTTTGAAGTGGATCATATTCTGCCCCTGTCAATCACCTTTGACGACTCGCTTGCGAACA

AGGTGCTCGTGTACGCAACGGCAAATCAGGAGAAAGGCCAGCGGACTCCGTATCAGGCG

CTCGACTCAATGGACGATGCGTGGTCATTCCGGGAGCTGAAGGCGTTCGTACGCGAGAG

CAAGACACTGAGCAACAAAAAGAAAGAGTATCTGCTGACAGAGGAGGACATCTCGAAAT

TCGATGTCAGGAAGAAGTTCATCGAGCGGAATCTTGTCGACACTCGCTACGCTTCCAGA

GTAGTACTGAACGCGCTCCAGGAACACTTTAGAGCGCACAAAATTGACACGAAGGTGTC

AGTGGTGAGAGGGCAGTTCACATCCCAACTCCGCCGACATTGGGGCATCGAAAAGACGC

GGGACACATATCACCATCATGCGGTGGACGCGCTGATTATTGCCGCTTCGTCCCAGTTG

AATCTCTGGAAAAAGCAGAAGAACACGCTGGTGTCGTATTCGGAGGATCAGCTTTTGGA

CATCGAAACCGGGGAGCTGATTTCCGACGATGAATACAAAGAATCGGTGTTTAAGGCAC

CATATCAGCATTTCGTGGACACGCTGAAGAGCAAAGAGTTTGAGGACAGCATCCTCTTT

TCGTACCAAGTGGACTCGAAGTTTAATCGCAAGATTTCAGACGCCACAATCTACGCGAC

GAGGCAGGCGAAGGTGGGCAAAGATAAAGCAGATGAAACCTACGTCCTTGGTAAAATCA

AGGACATCTACACTCAGGACGGGTACGATGCGTTCATGAAAATCTACAAGAAGGATAAG

TCGAAGTTTCTCATGTACCGCCACGATCCACAGACTTTCGAAAAAGTCATTGAGCCTAT

TTTGGAGAACTACCCTAACAAGCAAATCAACGAGAAAGGGAAAGAAGTCCCGTGCAACC

CCTTTCTGAAGTACAAGGAAGAGCACGGTTATATCCGCAAATACTCGAAGAAAGGAAAT

GGGCCTGAGATTAAGTCGCTTAAGTATTACGACTCAAAGTTGGGTAACCACATCGACAT

TACCCCGAAAGACTCCAACAACAAAGTCGTGTTGCAGTCCGTCTCGCCCTGGCGAGCAG

ATGTGTATTTTAATAAGACGACCGGCAAATATGAGATCCTTGGACTCAAATACGCAGAC

CTTCAATTCGAAAAGGGGACGGGCACTTATAAGATTTCACAAGAGAAGTACAACGACAT

CAAGAAAAAGGAAGGGGTCGATTCAGATTCGGAGTTCAAATTCACCCTCTACAAAAACG

ACCTCCTGCTTGTGAAGGACACAGAAACGAAGGAGCAGCAGCTCTTTCGGTTCCTCTCA

CGCACGATGCCCAAACAAAAACATTACGTCGAACTTAAACCTTACGATAAGCAAAAGTT
TGAAGGGGGAGAGGCACTGATCAAAGTATTGGGTAACGTAGCCAATAGCGGACAGTGTA
AGAAAGGGCTGGGAAAGTCCAATATCTCGATCTATAAAGTACGAACAGATGTATTGGGA
AACCAGCATATCATCAAAAATGAGGGGGATAAACCCAAACTCGATTTCAAGCGTCCTGC
TGCTACTAAGAAAGCTGGTCAAGCTAAGAAAAAGAAATAA

> U6-St_tracrRNA(7-97)
GAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGA
GATAATTGGAATTAATTTGACTGTAAACACAAAGATATTAGTACAAAATACGTGACGTA
GAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAATGGACTAT
CATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAA
AGGACGAAACACCGTTACTTAAATCTTGCAGAAGCTACAAAGATAAGGCTTCATGCCGA
AATCAACACCCTGTCATTTTATGGCAGGGTGTTTTCGTTATTTAA

>EMX1_TALEN_Left
ATGGACTATAAGGACCACGACGGAGACTACAAGGATCATGATATTGATTACAAAGACGA
TGACGATAAGATGGCCCCAAAGAAGAAGCGGAAGGTCGGTATCCACGGAGTCCCAGCAG
CCGTAGATTTGAGAACTTTGGGATATTCACAGCAGCAGCAGGAAAAGATCAAGCCCAAA
GTGAGGTCGACAGTCGCGCAGCATCACGAAGCGCTGGTGGGTCATGGGTTTACACATGC
CCACATCGTAGCCTTGTCGCAGCACCCTGCAGCCCTTGGCACGGTCGCCGTCAAGTACC
AGGACATGATTGCGGCGTTGCCGGAAGCCACACATGAGGCGATCGTCGGTGTGGGGAAA
CAGTGGAGCGGAGCCCGAGCGCTTGAGGCCCTGTTGACGGTCGCGGGAGAGCTGAGAGG
GCCTCCCCTTCAGCTGGACACGGGCCAGTTGCTGAAGATCGCGAAGCGGGGAGGAGTCA
CGGCGGTCGAGGCGGTGCACGCGTGGCGCAATGCGCTCACGGGAGCACCCCTCAACCTG
ACCCCAGAGCAGGTCGTGGCAATTGCGAGCAACCACGGGGGAAAGCAGGCACTCGAAAC
CGTCCAGAGGTTGCTGCCTGTGCTGTGCCAAGCGCACGGACTTACGCCAGAGCAGGTCG
TGGCAATTGCGAGCAACCACGGGGGAAAGCAGGCACTCGAAACCGTCCAGAGGTTGCTG
CCTGTGCTGTGCCAAGCGCACGGACTAACCCCAGAGCAGGTCGTGGCAATTGCGAGCAA

CATCGGGGGAAAGCAGGCACTCGAAACCGTCCAGAGGTTGCTGCCTGTGCTGTGCCAAG
CGCACGGGTTGACCCCAGAGCAGGTCGTGGCAATTGCGAGCAACCACGGGGGAAAGCAG
GCACTCGAAACCGTCCAGAGGTTGCTGCCTGTGCTGTGCCAAGCGCACGGCCTGACCCC
AGAGCAGGTCGTGGCAATTGCGAGCAACCACGGGGGAAAGCAGGCACTCGAAACCGTCC
AGAGGTTGCTGCCTGTGCTGTGCCAAGCGCACGGACTGACACCAGAGCAGGTCGTGGCA
ATTGCGAGCAACATCGGGGGAAAGCAGGCACTCGAAACCGTCCAGAGGTTGCTGCCTGT
GCTGTGCCAAGCGCACGGACTTACACCCGAACAAGTCGTGGCAATTGCGAGCAACCACG
GGGGAAAGCAGGCACTCGAAACCGTCCAGAGGTTGCTGCCTGTGCTGTGCCAAGCGCAC
GGACTTACGCCAGAGCAGGTCGTGGCAATTGCGAGCAACCACGGGGGAAAGCAGGCACT
CGAAACCGTCCAGAGGTTGCTGCCTGTGCTGTGCCAAGCGCACGGACTAACCCCAGAGC
AGGTCGTGGCAATTGCGAGCAACATCGGGGGAAAGCAGGCACTCGAAACCGTCCAGAGG
TTGCTGCCTGTGCTGTGCCAAGCGCACGGGTTGACCCCAGAGCAGGTCGTGGCAATTGC
GAGCAACATCGGGGGAAAGCAGGCACTCGAAACCGTCCAGAGGTTGCTGCCTGTGCTGT
GCCAAGCGCACGGCCTGACCCCAGAGCAGGTCGTGGCAATTGCGAGCAACCACGGGGGA
AAGCAGGCACTCGAAACCGTCCAGAGGTTGCTGCCTGTGCTGTGCCAAGCGCACGGACT
GACACCAGAGCAGGTCGTGGCAATTGCGAGCAACCACGGGGGAAAGCAGGCACTCGAAA
CCGTCCAGAGGTTGCTGCCTGTGCTGTGCCAAGCGCACGGACTCACGCCTGAGCAGGTA
GTGGCTATTGCATCCAACAACGGGGGCAGACCCGCACTGGAGTCAATCGTGGCCCAGCT
TTCGAGGCCGGACCCCGCGCTGGCCGCACTCACTAATGATCATCTTGTAGCGCTGGCCT
GCCTCGGCGGACGACCCGCCTTGGATGCGGTGAAGAAGGGGCTCCCGCACGCGCCTGCA
TTGATTAAGCGGACCAACAGAAGGATTCCCGAGAGGACATCACATCGAGTGGCAGGTTC
CCAACTCGTGAAGAGTGAACTTGAGGAGAAAAAGTCGGAGCTGCGGCACAAATTGAAAT
ACGTACCGCATGAATACATCGAACTTATCGAAATTGCTAGGAACTCGACTCAAGACAGA
ATCCTTGAGATGAAGGTAATGGAGTTCTTTATGAAGGTTTATGGATACCGAGGGAAGCA
TCTCGGTGGATCACGAAAACCCGACGGAGCAATCTATACGGTGGGGAGCCCGATTGATT
ACGGAGTGATCGTCGACACGAAAGCCTACAGCGGTGGGTACAATCTTCCCATCGGGCAG
GCAGATGAGATGCAACGTTATGTCGAAGAAAATCAGACCAGGAACAAACACATCAATCC
AAATGAGTGGTGGAAAGTGTATCCTTCATCAGTGACCGAGTTTAAGTTTTTGTTTGTCT

CTGGGCATTTCAAAGGCAACTATAAGGCCCAGCTCACACGGTTGAATCACATTACGAAC
TGCAATGGTGCGGTTTTGTCCGTAGAGGAACTGCTCATTGGTGGAGAAATGATCAAAGC
GGGAACTCTGACACTGGAAGAAGTCAGACGCAAGTTTAACAATGGCGAGATCAATTTCC
GCTCA

>EMX1_TALEN_Right

ATGGACTATAAGGACCACGACGGAGACTACAAGGATCATGATATTGATTACAAAGACGA
TGACGATAAGATGGCCCCAAAGAAGAAGCGGAAGGTCGGTATCCACGGAGTCCCAGCAG
CCGTAGATTTGAGAACTTTGGGATATTCACAGCAGCAGCAGGAAAAGATCAAGCCCAAA
GTGAGGTCGACAGTCGCGCAGCATCACGAAGCGCTGGTGGGTCATGGGTTTACACATGC
CCACATCGTAGCCTTGTCGCAGCACCCTGCAGCCCTTGGCACGGTCGCCGTCAAGTACC
AGGACATGATTGCGGCGTTGCCGGAAGCCACACATGAGGCGATCGTCGGTGTGGGGAAA
CAGTGGAGCGGAGCCCGAGCGCTTGAGGCCCTGTTGACGGTCGCGGGAGAGCTGAGAGG
GCCTCCCCTTCAGCTGGACACGGGCCAGTTGCTGAAGATCGCGAAGCGGGGAGGAGTCA
CGGCGGTCGAGGCGGTGCACGCGTGGCGCAATGCGCTCACGGGAGCACCCCTCAACCTG
ACCCCAGAGCAGGTCGTGGCAATTGCGAGCAACCACGGGGGAAAGCAGGCACTCGAAAC
CGTCCAGAGGTTGCTGCCTGTGCTGTGCCAAGCGCACGGACTTACGCCAGAGCAGGTCG
TGGCAATTGCGAGCAACCACGGGGGAAAGCAGGCACTCGAAACCGTCCAGAGGTTGCTG
CCTGTGCTGTGCCAAGCGCACGGACTAACCCCAGAGCAGGTCGTGGCAATTGCGAGCAA
CCACGGGGGAAAGCAGGCACTCGAAACCGTCCAGAGGTTGCTGCCTGTGCTGTGCCAAG
CGCACGGGTTGACCCCAGAGCAGGTCGTGGCAATTGCGAGCAACATCGGGGGAAAGCAG
GCACTCGAAACCGTCCAGAGGTTGCTGCCTGTGCTGTGCCAAGCGCACGGCCTGACCCC
AGAGCAGGTCGTGGCAATTGCGAGCAACCACGGGGGAAAGCAGGCACTCGAAACCGTCC
AGAGGTTGCTGCCTGTGCTGTGCCAAGCGCACGGACTGACACCAGAGCAGGTCGTGGCA
ATTGCGAGCCATGACGGGGGAAAGCAGGCACTCGAAACCGTCCAGAGGTTGCTGCCTGT

GCTGTGCCAAGCGCACGGACTTACACCCGAACAAGTCGTGGCAATTGCGAGCCATGACG

GGGGAAAGCAGGCACTCGAAACCGTCCAGAGGTTGCTGCCTGTGCTGTGCCAAGCGCAC

GGACTTACGCCAGAGCAGGTCGTGGCAATTGCGAGCCATGACGGGGGAAAGCAGGCACT

CGAAACCGTCCAGAGGTTGCTGCCTGTGCTGTGCCAAGCGCACGGACTAACCCCAGAGC

AGGTCGTGGCAATTGCGAGCAACGGAGGGGGAAAGCAGGCACTCGAAACCGTCCAGAGG

TTGCTGCCTGTGCTGTGCCAAGCGCACGGGTTGACCCCAGAGCAGGTCGTGGCAATTGC

GAGCAACGGAGGGGGAAAGCAGGCACTCGAAACCGTCCAGAGGTTGCTGCCTGTGCTGT

GCCAAGCGCACGGCCTGACCCCAGAGCAGGTCGTGGCAATTGCGAGCCATGACGGGGGA

AAGCAGGCACTCGAAACCGTCCAGAGGTTGCTGCCTGTGCTGTGCCAAGCGCACGGACT

GACACCAGAGCAGGTCGTGGCAATTGCGAGCAACGGAGGGGGAAAGCAGGCACTCGAAA

CCGTCCAGAGGTTGCTGCCTGTGCTGTGCCAAGCGCACGGACTCACGCCTGAGCAGGTA

GTGGCTATTGCATCCAACGGAGGGGGCAGACCCGCACTGGAGTCAATCGTGGCCCAGCT

TTCGAGGCCGGACCCCGCGCTGGCCGCACTCACTAATGATCATCTTGTAGCGCTGGCCT

GCCTCGGCGGACGACCCGCCTTGGATGCGGTGAAGAAGGGGCTCCCGCACGCGCCTGCA

TTGATTAAGCGGACCAACAGAAGGATTCCCGAGAGGACATCACATCGAGTGGCAGGTTC

CCAACTCGTGAAGAGTGAACTTGAGGAGAAAAAGTCGGAGCTGCGGCACAAATTGAAAT

ACGTACCGCATGAATACATCGAACTTATCGAAATTGCTAGGAACTCGACTCAAGACAGA

ATCCTTGAGATGAAGGTAATGGAGTTCTTTATGAAGGTTTATGGATACCGAGGGAAGCA

TCTCGGTGGATCACGAAAACCCGACGGAGCAATCTATACGGTGGGGAGCCCGATTGATT

ACGGAGTGATCGTCGACACGAAAGCCTACAGCGGTGGGTACAATCTTCCCATCGGGCAG

GCAGATGAGATGCAACGTTATGTCGAAGAAAATCAGACCAGGAACAAACACATCAATCC

AAATGAGTGGTGGAAAGTGTATCCTTCATCAGTGACCGAGTTTAAGTTTTTGTTTGTCT

CTGGGCATTTCAAAGGCAACTATAAGGCCCAGCTCACACGGTTGAATCACATTACGAAC

TGCAATGGTGCGGTTTTGTCCGTAGAGGAACTGCTCATTGGTGGAGAAATGATCAAAGC

GGGAACTCTGACACTGGAAGAAGTCAGACGCAAGTTTAACAATGGCGAGATCAATTTCC
GCTCA

References

1. P. A. Carr, G. M. Church, Genome engineering. *Nat Biotechnol* **27**, 1151 (Dec, 2009).
2. G. M. Church, Reading and writing omes. *Molecular systems biology* **9**, 642 (Jan 22, 2013).
3. R. R. Beerli, C. F. Barbas, 3rd, Engineering polydactyl zinc-finger transcription factors. *Nat Biotechnol* **20**, 135 (Feb, 2002).
4. M. R. Capecchi, Altering the genome by homologous recombination. *Science* **244**, 1288 (Jun 16, 1989).
5. K. M. Esvelt, H. H. Wang, Genome-scale engineering for systems and synthetic biology. *Molecular systems biology* **9**, 641 (Jan 22, 2013).
6. H. H. Wang *et al.*, Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894 (Aug 13, 2009).
7. H. H. Wang *et al.*, Genome-scale promoter engineering by coselection MAGE. *Nature methods* **9**, 591 (Jun, 2012).
8. H. J. Lee, E. Kim, J. S. Kim, Targeted chromosomal deletions in human cells using zinc finger nucleases. *Genome Res* **20**, 81 (Jan, 2010).
9. C. Sollu *et al.*, Autonomous zinc-finger nuclease pairs for targeted chromosomal deletion. *Nucleic Acids Res*, (Aug 16, 2010).
10. S. Thibodeau-Beganny, M. L. Maeder, J. K. Joung, Engineering single Cys2His2 zinc finger domains using a bacterial cell-based two-hybrid selection system. *Methods Mol Biol* **649**, 31 (2010).

11. S. Durai *et al.*, Zinc finger nucleases: custom-designed molecular scissors for genome engineering of plant and mammalian cells. *Nucleic Acids Res* **33**, 5978 (2005).

12. Y. G. Kim, J. Cha, S. Chandrasegaran, Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc Natl Acad Sci U S A* **93**, 1156 (Feb 6, 1996).

13. Y. G. Kim, S. Chandrasegaran, Chimeric restriction endonuclease. *Proc Natl Acad Sci U S A* **91**, 883 (Feb 1, 1994).

14. M. L. Maeder, S. Thibodeau-Beganny, J. D. Sander, D. F. Voytas, J. K. Joung, Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nat Protoc* **4**, 1471 (2009).

15. M. H. Porteus, D. Baltimore, Chimeric nucleases stimulate gene targeting in human cells. *Science* **300**, 763 (May 2, 2003).

16. M. H. Porteus, D. Carroll, Gene targeting using zinc finger nucleases. *Nat Biotechnol* **23**, 967 (Aug, 2005).

17. B. Gonzalez *et al.*, Modular system for the construction of zinc-finger libraries and proteins. *Nat Protoc* **5**, 791 (Apr, 2010).

18. C. O. Pabo, E. Peisach, R. A. Grant, Design and selection of novel Cys2His2 zinc finger proteins. *Annu Rev Biochem* **70**, 313 (2001).

19. J. D. Sander *et al.*, ZiFiT (Zinc Finger Targeter): an updated zinc finger engineering tool. *Nucleic Acids Res* **38 Suppl**, W462 (Jul 1, 2010).

20. J. Smith, J. M. Berg, S. Chandrasegaran, A detailed study of the substrate specificity of a chimeric restriction enzyme. *Nucleic Acids Res* **27**, 674 (Jan 15, 1999).

21. J. Smith *et al.*, A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. *Nucleic Acids Res* **34**, e149 (2006).

22. S. A. Wolfe, E. I. Ramm, C. O. Pabo, Combining structure-based design with phage display to create new Cys(2)His(2) zinc finger dimers. *Structure* **8**, 739 (Jul 15, 2000).

23. J. Boch *et al.*, Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**, 1509 (Dec 11, 2009).

24. J. Boch, U. Bonas, Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annu Rev Phytopathol* **48**, 419 (Sep 8, 2010).

25. M. J. Moscou, A. J. Bogdanove, A simple cipher governs DNA recognition by TAL effectors. *Science* **326**, 1501 (Dec 11, 2009).

26. A. J. Bogdanove, S. Schornack, T. Lahaye, TAL effectors: finding plant genes for disease and defense. *Curr Opin Plant Biol* **13**, 394 (Aug, 2010).

27. J. D. Sander *et al.*, Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). *Nature methods* **8**, 67 (Jan, 2011).

28. N. E. Sanjana *et al.*, A transcription activator-like effector toolbox for genome engineering. *Nature protocols* **7**, 171 (Jan, 2012).

29. L. Cong, R. Zhou, Y. C. Kuo, M. Cunniff, F. Zhang, Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nature communications* **3**, 968 (2012).

30. J. E. Garneau *et al.*, The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67 (Nov 4, 2010).

31. H. Deveau, J. E. Garneau, S. Moineau, CRISPR/Cas system and its role in phage-bacteria interactions. *Annual review of microbiology* **64**, 475 (2010).

32. P. Horvath, R. Barrangou, CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167 (Jan 8, 2010).

33. L. Cong *et al.*, Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819 (Feb 15, 2013).

34. G. E. Meister, S. Chandrasegaran, M. Ostermeier, Heterodimeric DNA methyltransferases as a platform for creating designer zinc finger methyltransferases for targeted DNA methylation in cells. *Nucleic Acids Res* **38**, 1749 (Mar, 2010).

35. P. Blancafort, L. Magnenat, C. F. Barbas, 3rd, Scanning the human genome with combinatorial transcription factor libraries. *Nat Biotechnol* **21**, 269 (Mar, 2003).

36. L. E. Rosen *et al.*, Homing endonuclease I-CreI derivatives with novel DNA target specificities. *Nucleic Acids Res* **34**, 4791 (2006).

37. S. Grizot *et al.*, Efficient targeting of a SCID gene by an engineered single-chain homing endonuclease. *Nucleic Acids Res* **37**, 5405 (Sep, 2009).

38. J. C. Miller *et al.*, A TALE nuclease architecture for efficient genome editing. *Nat Biotech* **advance online publication**, (2010).

39. M. Christian *et al.*, Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186**, 757 (Oct, 2010).

40. R. Morbitzer, P. Romer, J. Boch, T. Lahaye, Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors. *Proc Natl Acad Sci U S A*, (Nov 24, 2010).

41. S. Kay, S. Hahn, E. Marois, R. Wieduwild, U. Bonas, Detailed analysis of the DNA recognition motifs of the Xanthomonas type III effectors AvrBs3 and AvrBs3Deltarep16. *Plant J* **59**, 859 (Sep, 2009).

42. P. Romer *et al.*, Recognition of AvrBs3-like proteins is mediated by specific binding to promoters of matching pepper Bs3 alleles. *Plant Physiol* **150**, 1697 (Aug, 2009).

43. C. Engler, R. Gruetzner, R. Kandzia, S. Marillonnet, Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS One* **4**, e5553 (2009).

44. C. Engler, R. Kandzia, S. Marillonnet, A one pot, one step, precision cloning method with high throughput capability. *PLoS One* **3**, e3647 (2008).

45. S. Kay, S. Hahn, E. Marois, G. Hause, U. Bonas, A bacterial effector acts as a plant transcription factor and induces a cell size regulator. *Science* **318**, 648 (Oct 26, 2007).

46. P. Romer *et al.*, Plant pathogen recognition mediated by promoter activation of the pepper Bs3 resistance gene. *Science* **318**, 645 (Oct 26, 2007).

47. H. Scholze, J. Boch, TAL effector-DNA specificity. *Virulence* **1**, 428 (Dec 22, 2010).

48. R. R. Beerli, D. J. Segal, B. Dreier, C. F. Barbas, 3rd, Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using

polydactyl zinc finger proteins constructed from modular building blocks. *Proc Natl Acad Sci U S A* **95**, 14628 (Dec 8, 1998).

49. N. Xu, T. Papagiannakopoulos, G. Pan, J. A. Thomson, K. S. Kosik, MicroRNA-145 regulates OCT4, SOX2, and KLF4 and represses pluripotency in human embryonic stem cells. *Cell* **137**, 647 (May 15, 2009).

50. Z. Wei *et al.*, Klf4 interacts directly with Oct4 and Sox2 to promote reprogramming. *Stem Cells* **27**, 2969 (Dec, 2009).

51. R. M. Gordley, C. A. Gersbach, C. F. Barbas, 3rd, Synthesis of programmable integrases. *Proc Natl Acad Sci U S A* **106**, 5053 (Mar 31, 2009).

52. F. Zhang *et al.*, Multimodal fast optical interrogation of neural circuitry. *Nature* **446**, 633 (Apr 5, 2007).

53. J. C. Miller *et al.*, A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **29**, 143 (Feb, 2011).

54. R. Morbitzer, P. Romer, J. Boch, T. Lahaye, Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors. *Proc Natl Acad Sci U S A* **107**, 21617 (Dec 14, 2010).

55. E. Weber, R. Gruetzner, S. Werner, C. Engler, S. Marillonnet, Assembly of Designer TAL Effectors by Golden Gate Cloning. *PloS one* **6**, e19722 (2011).

56. T. Cermak *et al.*, Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic acids research* **39**, e82 (Jul 1, 2011).

57. R. Geiβler *et al.*, Transcriptional Activators of Human Genes with Programmable DNA-Specificity. *PloS one* **6**, e19509 (2011).

58. T. Li *et al.*, Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. *Nucleic acids research*, (Mar 31, 2011).

59. R. Morbitzer, J. Elsaesser, J. Hausner, T. Lahaye, Assembly of custom TALE-type DNA binding domains by modular cloning. *Nucleic acids research*, (Mar 18, 2011).

60. A. J. Wood *et al.*, Targeted genome editing across species using ZFNs and TALENs. *Science* **333**, 307 (Jul 15, 2011).

61. D. Hockemeyer *et al.*, Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology*, (Jul 7, 2011).

62. T. Li *et al.*, TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic acids research* **39**, 359 (Jan 1, 2011).

63. M. M. Mahfouz *et al.*, De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 2623 (Feb 8, 2011).

64. J. Boch, U. Bonas, Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annu Rev Phytopathol* **48**, 419 (Sep 8, 2010).

65. A. J. Bogdanove, S. Schornack, T. Lahaye, TAL effectors: finding plant genes for disease and defense. *Curr Opin Plant Biol* **13**, 394 (Aug, 2010).

66. S. Kay, S. Hahn, E. Marois, R. Wieduwild, U. Bonas, Detailed analysis of the DNA recognition motifs of the Xanthomonas type III effectors AvrBs3 and AvrBs3Deltarep16. *Plant J* **59**, 859 (Sep, 2009).

67. P. Romer *et al.*, Recognition of AvrBs3-like proteins is mediated by specific binding to promoters of matching pepper Bs3 alleles. *Plant Physiol* **150**, 1697 (Aug, 2009).

68. A. Hinnen, J. B. Hicks, G. R. Fink, Transformation of yeast. *Proceedings of the National Academy of Sciences of the United States of America* **75**, 1929 (Apr, 1978).

69. J. W. Szostak, T. L. Orr-Weaver, R. J. Rothstein, F. W. Stahl, The double-strand-break repair model for recombination. *Cell* **33**, 25 (May, 1983).

70. K. R. Thomas, K. R. Folger, M. R. Capecchi, High frequency targeting of genes to specific sites in the mammalian genome. *Cell* **44**, 419 (Feb 14, 1986).

71. Z. Ivics, P. B. Hackett, R. H. Plasterk, Z. Izsvak, Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* **91**, 501 (Nov 14, 1997).

72. K. Kawakami, A. Shima, N. Kawakami, Identification of a functional transposase of the Tol2 element, an Ac-like element from the Japanese medaka fish, and its transposition in the zebrafish germ lineage. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 11403 (Oct 10, 2000).

73. K. Akagi *et al.*, Cre-mediated somatic site-specific recombination in mice. *Nucleic acids research* **25**, 1766 (May 1, 1997).

74. J. C. Epinat *et al.*, A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucleic acids research* **31**, 2952 (Jun 1, 2003).

75. C. Lois, E. J. Hong, S. Pease, E. J. Brown, D. Baltimore, Germline transmission and tissue-specific expression of transgenes delivered by lentiviral vectors. *Science* **295**, 868 (Feb 1, 2002).

76. I. F. Khan, R. K. Hirata, D. W. Russell, AAV-mediated gene targeting methods for human cells. *Nature protocols* **6**, 482 (Apr, 2011).

77. N. P. Pavletich, C. O. Pabo, Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* **252**, 809 (May 10, 1991).

78. A. Klug, The discovery of zinc fingers and their development for practical applications in gene regulation and genome manipulation. *Quarterly reviews of biophysics* **43**, 1 (Feb, 2010).

79. M. L. Maeder, S. Thibodeau-Beganny, J. D. Sander, D. F. Voytas, J. K. Joung, Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nat Protoc* **4**, 1471 (2009).

80. J. S. Kim, H. J. Lee, D. Carroll, Genome editing with modularly assembled zinc-finger nucleases. *Nature methods* **7**, 91; author reply 91 (Feb, 2010).

81. E. E. Perez *et al.*, Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature biotechnology* **26**, 808 (Jul, 2008).

82. R. A. Keenholtz, S. J. Rowland, M. R. Boocock, W. M. Stark, P. A. Rice, Structural basis for catalytic activation of a serine recombinase. *Structure* **19**, 799 (Jun 8, 2011).

83. C. A. Gersbach, T. Gaj, R. M. Gordley, A. C. Mercer, C. F. Barbas, 3rd, Targeted plasmid integration into the human genome by an engineered zinc-finger recombinase. *Nucleic acids research*, (Jun 7, 2011).

84. T. Gaj, A. C. Mercer, C. A. Gersbach, R. M. Gordley, C. F. Barbas, 3rd, Structure-guided reprogramming of serine recombinase DNA sequence specificity. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 498 (Jan 11, 2011).

85. F. D. Urnov *et al.*, Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435**, 646 (Jun 2, 2005).

86. M. H. Wilson, J. M. Kaminski, A. L. George, Jr., Functional zinc finger/sleeping beauty transposase chimeras exhibit attenuated overproduction inhibition. *FEBS letters* **579**, 6205 (Nov 7, 2005).

87. P. Huang *et al.*, Heritable gene targeting in zebrafish using customized TALENs. *Nature biotechnology* **29**, 699 (2011).

88. J. D. Sander *et al.*, Targeted gene disruption in somatic zebrafish cells using engineered TALENs. *Nature biotechnology* **29**, 697 (2011).

89. L. Tesson *et al.*, Knockout rats generated by embryo microinjection of TALENs. *Nature biotechnology* **29**, 695 (2011).

90. E. Weber, C. Engler, R. Gruetzner, S. Werner, S. Marillonnet, A modular cloning system for standardized assembly of multigene constructs. *PloS one* **6**, e16765 (2011).

91. P. Huertas, DNA resection in eukaryotes: deciding how to fix the break. *Nature structural & molecular biology* **17**, 11 (Jan, 2010).

92. T. Nolan, R. E. Hands, S. A. Bustin, Quantification of mRNA using real-time RT-PCR. *Nature protocols* **1**, 1559 (2006).

93. D. Y. Guschin *et al.*, A rapid and general assay for monitoring endogenous gene modification. *Methods in molecular biology* **649**, 247 (2010).

94. F. Zhang *et al.*, High frequency targeted mutagenesis in Arabidopsis thaliana using zinc finger nucleases. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 12028 (Jun 29, 2010).

95. A. A. Buzdin, in *Nucleic Acids Hybridization,* A. Buzdin, Lukyanov, S., Ed. (Springer, 2007), pp. 211-239.

96. B. J. Till, C. Burtner, L. Comai, S. Henikoff, Mismatch cleavage by single-strand specific nucleases. *Nucleic acids research* **32**, 2632 (2004).

97. J. J. Babon, M. McKenzie, R. G. Cotton, The use of resolvases T4 endonuclease VII and T7 endonuclease I in mutation detection. *Molecular biotechnology* **23**, 73 (Jan, 2003).

98. B. Yang *et al.*, Purification, cloning, and characterization of the CEL I nuclease. *Biochemistry* **39**, 3533 (Apr 4, 2000).

99. J. Kulinski, D. Besack, C. A. Oleykowski, A. K. Godwin, A. T. Yeung, CEL I enzymatic mutation detection assay. *BioTechniques* **29**, 44 (Jul, 2000).

100. C. A. Oleykowski, C. R. Bronson Mullins, A. K. Godwin, A. T. Yeung, Mutation detection using a novel plant endonuclease. *Nucleic acids research* **26**, 4597 (Oct 15, 1998).

101. M. W. Pfaffl, A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic acids research* **29**, e45 (May 1, 2001).

102. M. T. Murakami *et al.*, The repeat domain of the type III effector protein PthA shows a TPR-like structure and undergoes conformational changes upon DNA interaction. *Proteins* **78**, 3386 (Dec, 2010).

103. H. Scholze, J. Boch, TAL effectors are remote controls for gene activation. *Current opinion in microbiology* **14**, 47 (Feb, 2011).

104. M. M. Mahfouz *et al.*, Targeted transcriptional repression using a chimeric TALE-SRDX repressor protein. *Plant molecular biology* **78**, 311 (Feb, 2012).

105. A. J. Bogdanove, D. F. Voytas, TAL effectors: customizable proteins for DNA targeting. *Science* **333**, 1843 (Sep 30, 2011).

106. D. E. Ayer, C. D. Laherty, Q. A. Lawrence, A. P. Armstrong, R. N. Eisenman, Mad proteins contain a dominant transcription repression domain. *Molecular and cellular biology* **16**, 5772 (Oct, 1996).

107. A. N. Mak, P. Bradley, R. A. Cernadas, A. J. Bogdanove, B. L. Stoddard, The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **335**, 716 (Feb 10, 2012).

108. D. Deng *et al.*, Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* **335**, 720 (Feb 10, 2012).

109. C. Batchelder *et al.*, Transcriptional repression by the Caenorhabditis elegans germ-line protein PIE-1. *Genes & development* **13**, 202 (Jan 15, 1999).

110. E. Tour, C. T. Hittinger, W. McGinnis, Evolutionarily conserved domains required for activation and repression functions of the Drosophila Hox protein Ultrabithorax. *Development* **132**, 5271 (Dec, 2005).

111. S. B. Tiwari, G. Hagen, T. J. Guilfoyle, Aux/IAA proteins contain a potent transcriptional repression domain. *The Plant cell* **16**, 533 (Feb, 2004).

112. J. F. Margolin *et al.*, Kruppel-associated boxes are potent transcriptional repression domains. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 4509 (May 10, 1994).

113. M. Almlof, J. Aqvist, A. O. Smalas, B. O. Brandsdal, Probing the effect of point mutations at protein-protein interfaces with free energy calculations. *Biophysical journal* **90**, 433 (Jan 15, 2006).

114. J. Wang, Y. Deng, B. Roux, Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophysical journal* **91**, 2798 (Oct 15, 2006).

115. R. Zhou, P. Das, A. K. Royyuru, Single mutation induced H3N2 hemagglutinin antibody neutralization: a free energy perturbation study. *The journal of physical chemistry. B* **112**, 15813 (Dec 11, 2008).

116. J. D. Chodera *et al.*, Alchemical free energy methods for drug discovery: progress and challenges. *Current opinion in structural biology* **21**, 150 (Apr, 2011).

117. J. C. Miller *et al.*, An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature biotechnology* **25**, 778 (Jul, 2007).

118. D. Reyon *et al.*, FLASH assembly of TALENs for high-throughput genome editing. *Nature biotechnology* **30**, 460 (May, 2012).

119. B. L. Stoddard, Homing endonuclease structure and function. *Quarterly reviews of biophysics* **38**, 49 (Feb, 2005).

120. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816 (Aug 17, 2012).

121. G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E2579 (Sep 25, 2012).

122. K. S. Makarova *et al.*, Evolution and classification of the CRISPR-Cas systems. *Nature reviews. Microbiology* **9**, 467 (Jun, 2011).

123. D. Bhaya, M. Davison, R. Barrangou, CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annual review of genetics* **45**, 273 (2011).

124. E. Deltcheva *et al.*, CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602 (Mar 31, 2011).

125. R. Sapranauskas *et al.*, The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic acids research* **39**, 9275 (Nov, 2011).

126. A. H. Magadan, M. E. Dupuis, M. Villion, S. Moineau, Cleavage of phage DNA by the Streptococcus thermophilus CRISPR3-Cas system. *PloS one* **7**, e40913 (2012).

127. H. Deveau *et al.*, Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *Journal of bacteriology* **190**, 1390 (Feb, 2008).

128. F. J. Mojica, C. Diez-Villasenor, J. Garcia-Martinez, C. Almendros, Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733 (Mar, 2009).

129. M. Jinek, J. A. Doudna, A three-dimensional view of the molecular machinery of RNA interference. *Nature* **457**, 405 (Jan 22, 2009).

130. C. D. Malone, G. J. Hannon, Small RNAs as guardians of the genome. *Cell* **136**, 656 (Feb 20, 2009).

131. G. Meister, T. Tuschl, Mechanisms of gene silencing by double-stranded RNA. *Nature* **431**, 343 (Sep 16, 2004).

132. M. T. Certo *et al.*, Tracking genome engineering outcome at individual DNA breakpoints. *Nature methods* **8**, 671 (Aug, 2011).

133. Mali et al.

134. P. A. Carr, G. M. Church, Genome engineering. *Nat Biotechnol* **27**, 1151 (Dec, 2009).

135. P. J. Belshaw, S. N. Ho, G. R. Crabtree, S. L. Schreiber, Controlling protein association and subcellular localization with a synthetic ligand that induces heterodimerization of proteins. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 4604 (May 14, 1996).

136. S. Crosson, S. Rajagopal, K. Moffat, The LOV domain family: photoresponsive signaling modules coupled to diverse output domains. *Biochemistry* **42**, 2 (Jan 14, 2003).

137. S. R. Cutler, P. L. Rodriguez, R. R. Finkelstein, S. R. Abrams, Abscisic acid: emergence of a core signaling network. *Annual review of plant biology* **61**, 651 (2010).

138. A. Fegan, B. White, J. C. Carlson, C. R. Wagner, Chemically controlled protein assembly: techniques and applications. *Chemical reviews* **110**, 3315 (Jun 9, 2010).

139. S. N. Ho, S. R. Biggar, D. M. Spencer, S. L. Schreiber, G. R. Crabtree, Dimeric ligands define a role for transcriptional activation domains in reinitiation. *Nature* **382**, 822 (Aug 29, 1996).

140. M. J. Kennedy *et al.*, Rapid blue-light-mediated induction of protein interactions in living cells. *Nature methods* **7**, 973 (Dec, 2010).

141. D. Strickland, K. Moffat, T. R. Sosnick, Light-activated DNA binding in a designed allosteric protein. *Proc Natl Acad Sci U S A* **105**, 10709 (Aug 5, 2008).

142. D. Strickland *et al.*, Rationally improving LOV domain-based photoswitches. *Nat Methods* **7**, 623 (Aug, 2010).

143. B. D. Zoltowski, B. Vaccaro, B. R. Crane, Mechanism-based tuning of a LOV domain photoreceptor. *Nat Chem Biol* **5**, 827 (Nov, 2009).

144. J. Zuo, N. H. Chua, Chemical-inducible systems for regulated expression of plant genes. *Current opinion in biotechnology* **11**, 146 (Apr, 2000).

145. G. Stephanopoulos, Synthetic Biology and Metabolic Engineering. *Acs Synthetic Biology* **1**, 514 (Nov, 2012).

146. V. G. Yadav, M. De Mey, C. G. Lim, P. K. Ajikumar, G. Stephanopoulos, The future of metabolic engineering and synthetic biology: Towards a systematic practice. *Metabolic Engineering* **14**, 233 (May, 2012).

147. A. A. Cheng, T. K. Lu, Synthetic biology: an emerging engineering discipline. *Annual review of biomedical engineering* **14**, 155 (2012).

148. J. D. Keasling, Synthetic biology and the development of tools for metabolic engineering. *Metabolic engineering* **14**, 189 (May, 2012).

149. H. H. Wang, G. M. Church, Multiplexed genome engineering and genotyping methods applications for synthetic biology and metabolic engineering. *Methods in enzymology* **498**, 409 (2011).

150. H. Gao, X. Wu, J. Chai, Z. Han, Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell research* **22**, 1716 (Dec, 2012).

151. Y. Kim *et al.*, A library of TAL effector nucleases spanning the human genome. *Nature biotechnology*, (Feb 17, 2013).

152. T. Gaj, A. C. Mercer, S. J. Sirk, H. L. Smith, C. F. Barbas, 3rd, A comprehensive approach to zinc-finger recombinase customization enables genomic targeting in human cells. *Nucleic acids research*, (Feb 7, 2013).

153. A. C. Mercer, T. Gaj, R. P. Fuller, C. F. Barbas, 3rd, Chimeric TALE recombinases with programmable DNA sequence specificity. *Nucleic acids research* **40**, 11163 (Nov, 2012).

154. I. H. Chou, T. Chouard, Neuropsychiatric disease. *Nature* **455**, 889 (Oct 16, 2008).

155. H. U. Wittchen, The size and burden of mental disorders in Europe - an ECNP task force report. *European Neuropsychopharmacology* **15**, S313 (Oct, 2005).

156. H. U. Wittchen, F. Jacobi, Size and burden of mental disorders in Europe - a critical review and appraisal of 27 studies. *European Neuropsychopharmacology* **15**, 357 (Aug, 2005).

157. H. U. Wittchen *et al.*, The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European Neuropsychopharmacology* **21**, 655 (Sep, 2011).

158. G. R. Uhl, R. W. Grow, The burden of complex genetics in brain disorders. *Archives of General Psychiatry* **61**, 223 (Mar, 2004).

159. P. J. Ross, J. Ellis, Modeling complex neuropsychiatric disease with induced pluripotent stem cells. *F1000 Biol Rep* **2**, 84 (2010).

160. M. Burmeister, M. G. McInnis, S. Zollner, Psychiatric genetics: progress amid controversy. *Nat Rev Genet* **9**, 527 (Jul, 2008).

161. E. H. Cook, Jr., S. W. Scherer, Copy-number variations associated with neuropsychiatric conditions. *Nature* **455**, 919 (Oct 16, 2008).

162. V. M. Bedell *et al.*, In vivo genome editing using a high-efficiency TALEN system. *Nature* **491**, 114 (Nov 1, 2012).

163. D. F. Carlson *et al.*, Efficient TALEN-mediated gene knockout in livestock. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 17382 (Oct 23, 2012).

164. Y. Lei *et al.*, Efficient targeted gene disruption in Xenopus embryos using engineered transcription activator-like effector nucleases (TALENs). *Proceedings of the National Academy of Sciences of the United States of America* **109**, 17484 (Oct 23, 2012).

165. N. Sun, J. Liang, Z. Abil, H. Zhao, Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. *Molecular bioSystems* **8**, 1255 (Apr, 2012).

166. K. Takahashi, S. Yamanaka, Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663 (Aug 25, 2006).

167. G. Lee, L. Studer, Induced pluripotent stem cell technology for the study of human disease. *Nat Methods* **7**, 25 (Jan, 2010).

168. M. C. Marchetto, B. Winner, F. H. Gage, Pluripotent stem cells in neurodegenerative and neurodevelopmental diseases. *Hum Mol Genet* **19**, R71 (Apr 15, 2010).

169. K. Kim *et al.*, Epigenetic memory in induced pluripotent stem cells. *Nature* **467**, 285 (Sep 16, 2010).

170. I. H. Park *et al.*, Disease-specific induced pluripotent stem cells. *Cell* **134**, 877 (Sep 5, 2008).

171. D. Hockemeyer *et al.*, Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology* **29**, 731 (Aug, 2011).

172. N. Holt *et al.*, Human hematopoietic stem/progenitor cells modified by zinc-finger nucleases targeted to CCR5 control HIV-1 in vivo. *Nature biotechnology* **28**, 839 (Aug, 2010).

173. R. Tomanin, M. Scarpa, Why do we need new gene therapy viral vectors? Characteristics, limitations and future perspectives of viral vector transduction. *Current gene therapy* **4**, 357 (Dec, 2004).

174. M. Cavazzana-Calvo, A. Thrasher, F. Mavilio, The future of gene therapy. *Nature* **427**, 779 (Feb 26, 2004).

175. P. D. Robbins, S. C. Ghivizzani, Viral vectors for gene therapy. *Pharmacology & therapeutics* **80**, 35 (Oct, 1998).

176. M. I. Phillips, Gene, stem cell, and future therapies for orphan diseases. *Clinical pharmacology and therapeutics* **92**, 182 (Aug, 2012).

177. P. Perez-Pinera, D. G. Ousterout, C. A. Gersbach, Advances in targeted genome editing. *Current opinion in chemical biology* **16**, 268 (Aug, 2012).

178. O. Humbert, L. Davis, N. Maizels, Targeted gene therapies: tools, applications, optimization. *Critical reviews in biochemistry and molecular biology* **47**, 264 (May-Jun, 2012).

179. C. Sheridan, Gene therapy finds its niche. *Nature biotechnology* **29**, 121 (Feb, 2011).

180. P. A. LeWitt *et al.*, AAV2-GAD gene therapy for advanced Parkinson's disease: a double-blind, sham-surgery controlled, randomised trial. *Lancet neurology* **10**, 309 (Apr, 2011).

181. F. Ferrua, I. Brigida, A. Aiuti, Update on gene therapy for adenosine deaminase-deficient severe combined immunodeficiency. *Current opinion in allergy and clinical immunology* **10**, 551 (Dec, 2010).

182. N. Cartier, P. Aubourg, Hematopoietic stem cell transplantation and hematopoietic stem cell gene therapy in X-linked adrenoleukodystrophy. *Brain pathology* **20**, 857 (Jul, 2010).

183. C. Zuber *et al.*, Delivery of single-chain antibodies (scFvs) directed against the 37/67 kDa laminin receptor into mice via recombinant adeno-associated viral

vectors for prion disease gene therapy. *The Journal of general virology* **89**, 2055 (Aug, 2008).

184. J. W. Bainbridge *et al.*, Effect of gene therapy on visual function in Leber's congenital amaurosis. *The New England journal of medicine* **358**, 2231 (May 22, 2008).

185. C. L. Cepko, Emerging gene therapies for retinal degenerations. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **32**, 6415 (May 9, 2012).

186. S. Kosuri *et al.*, Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat Biotechnol* **28**, 1295 (Dec, 2010).

187. R. Barrangou, P. Horvath, CRISPR: new horizons in phage resistance and strain identification. *Annual review of food science and technology* **3**, 143 (2012).

188. S. J. Brouns *et al.*, Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960 (Aug 15, 2008).